DÉVELOPPEMENT D'OUTILS BIO-INFORMATIQUES POUR L'ANALYSE DE DONNÉES ÉPIGÉNOMIQUES AVEC RÉFÉRENCE EXTERNE ET POUR L'ÉVALUATION DU NOMBRE DE COUPLES À RISQUE À PARTIR DE FICHIERS DE VARIANTS GÉNÉTIQUES

par

Marc-Antoine Robert

mémoire présenté au Département de biologie en vue
de l'obtention du grade de maître ès sciences (M. Sc.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, Juillet 2018

Le 16 Juillet 2018

le jury a accepté le mémoire de Monsieur Marc-Antoine Robert
dans sa version finale.

Membres du Jury

Professeur Pierre-Étienne Jacques
Directeur de recherche
Département de Biologie

Professeur Michelle Scott
Évaluateur interne
Département de Biochimie

Professeur Luc Gaudreau
Président-rapporteur
Département de Biologie

SOMMAIRE

L'augmentation de l'accessibilité des techniques de séquençages à haut débit, a permis de générer de grands nombres de donnés avec une précision au nucléotide près. Ces données découlent de plusieurs types d'expériences génomiques et l'évolution de ces expériences demande une toute aussi rapide implémentation d'outils informatiques faciles d'utilisation pour les biologistes. Ce mémoire présente deux outils indépendants et leurs applications dans deux domaines de la biologie : la génomique et la génétique.

Dans le cadre du projet en génomique, l'outil de normalisation SpkNorm a été développé dans le but d'analyser des données de localisation de la chromatine avec référence externe. Basés sur ces résultats et sur d'autres expériences effectuées, nous avons montré que contrairement à ce qui est retrouvé chez *Saccharomyces cerevisiae*, la terminaison de la transcription par l'ARN polymérase II des gènes non-codants chez *Schizosaccharomyces pombe* est causée par le modèle *Torpedo*. En effet, chez *S. cerevisiae* ce mécanisme est responsable de la terminaison des gènes codants pour des protéines, alors que la terminaison de la transcription des gènes non-codants est prise en charge par le complexe Nrd1-Nab3-Seb1.

Dans le cadre du projet en génétique, nous avons développé l'outil modulaire et paramétrable MockScreen, servant à identifier la fréquence de nombre de couples à risque d'engendrer un enfant ayant une maladie récessive rare. Afin d'amener des pistes de réponses sur l'aspect économique de séquencer systématiquement tous les futurs parents, avoir une méthode précise pour identifier le nombre de couples à risque est primordial. Plusieurs dizaines de milliers d'individus provenant de diverses sources sont nécessaires pour une telle analyse. Par contre, l'accès à ces données et la

compatibilité entre ces données (générées avec différents logiciels ou différentes versions) furent beaucoup plus complexes qu'attendu. Par conséquent, cette étude se concentre sur un petit nombre d'individus (~2000) et ne fait qu'une preuve de concept. Bien que la fréquence théorique de couples à risque dans une population puisse être calculée facilement en prenant les données de fréquences populationnelles des variants par gène, nos résultats suggèrent que la génération de couples synthétiques est essentielle pour considérer le génotype complet des individus et ainsi éviter un biais.

Mots-clés : outil Bio-informatique, génomique, terminaison de transcription, *Schizosaccharomyces pombe,* ChIP-Seq, normalisation, génétique, variant, maladie mendélienne et fréquence allélique

REMERCIEMENTS

La réalisation de ce mémoire a été possible grâce à la participation directe ou indirecte de plusieurs personnes à qui je voudrais témoigner toute ma reconnaissance. Avant tout, les plus grands honneurs vont au directeur de ce mémoire, Pierre-Étienne Jacques sans qui je n'aurais pas eu la chance et les aptitudes de m'investir dans un tel projet. En effet, son mentorat a permis d'alimenter ma réflexion, de bâtir mes compétences et de me développer une véritable passion professionnelle.

J'aimerais aussi remercier François Bachand, ainsi que Sébastien Lévesque pour leur implication directe dans le projet et d'avoir eu confiance en moi pour avoir participé à un de leur projet de recherche respectif. Leurs conseils, ainsi que ceux de Sébastien Rodrigue, ont permis l'aboutissement de ce mémoire.

Il impossible de mettre sur pied de telle recherche sans financement. Pour cette raison, un grand merci au CRSNG, au FRQS et l'Université de Sherbrooke d'avoir financé cette recherche.

Pour finir, je voulais exprimer ma reconnaissance envers toutes les personnes qui m'entoure professionnellement et personnellement. Leur support moral non négligeable a permis de traverser les moments plus difficiles et de mettre à terme ce mémoire. Merci à mon conjoint Frédérick Lisso, mes parents, ma famille, mes collègues et mes amis !

LISTE DES ABRÉVIATIONS

ARN	Acide ribonucléique
3'READS	*3' region extraction and deep sequencing*
ADN	Acide désoxyribonucléique
ARNlnc	Long ARN non codant
ARNm	ARN messager
ARNnc	ARN non-codant
ARNpol	ARN polymérase
ARNr	ARN ribosomal
ARNsn	Petit ARN nucléaire
ARNsno	Petit ARN nucléolaire
ARNt	ARN de transfert
ChIP-chip	Immunoprécipitation de chromatine sur *chip*
ChIP-Seq	Immunoprécipitation de chromatine couplé au séquençage
ChIP-Seq-SI	Immunoprécipitation de chromatine couplé au séquençage avec référence externe
CID	Domaine d'interaction du CTD
CRCHUS	Centre de Recherche du Centre Hospitalier de l'Université de Sherbrooke
CTD	Domaine carboxyle terminal
CUT	Transcrit cryptique instable
e.g.	*Exempli gratia* (par exemple)
ExAC	*Exome Aggregate Consortium*
gnomAD	*Genome Aggregation Database*
HH	Couple dont chaque individu a au moins un variant d'impact élevé dans le même gène
HHM	Union des types de couple HH et HM

HM	Couple dont un individu avec au moins un variant d'impact élevé et l'autre avec au moins un variant d'impact moyen dans le même gène
IP	Ensemble de données issues de l'immunoprécipitation
kb	Kilobase
M	Million
MACS	*Model-based analysis of ChIP-Seq*
MM	Couple ayant pour chaque individu au moins un variant d'impact moyen dans le même gène.
Mut	Souche Mutante
N	Nucleotide non-déterminé
NCIS	*Normalization on ChIP-Seq*
NIH	*National Institutes of Health*
NNS	Nrd1-Nab3-Seb1
pb	Paire de bases
Pro	Proline
qPCR	Réaction en chaine de polymérase quantitatif
Queue polyA	Queue polyadénylé
RPKM	*Read per Kilobase per Million*
S. cerevisiae	*Saccharomyces cerevisiae*
S. pombe	*Schizosaccharomyces pombe*
Ser	Sérine
Ser2P	Forme phosphorylée de la serine 2 du CTD de l'ARNpol II
Ser5P	Forme phosphorylée de la serine 5 du CTD de l'ARNpol II
Ser7P	Forme phosphorylée de la serine 7 du CTD de l'ARNpol II
SI	Référence externe, *spike-in*
Site polyA	Site de polyadénylation
SNP	Polymorphisme d'un seul nucléotide
Thr	Thréonine
Tyr	Tyrosine
Tyr1P	Forme phosphorylée de la tyrosine 1 du CTD de l'ARNpol II

VCF	*Variant call format*
WES	Séquençage complet de l'exome
WT	Souche sauvage

LISTE DES TABLEAUX

LISTE DES FIGURES

CHAPITRE1
INTRODUCTION GÉNÉRALE

Le développement des technologies de séquençage au cours des dernières années a permis de générer des quantités massives de données biologiques et nécessite de plus en plus une approche multidisciplinaire biologie/informatique lors de publications. La génomique et la génétique font partie des moteurs principaux de cette nouvelle ère. En effet, l'avancement de leurs modèles expérimentaux demande un constant remaniement des analyses post-expérimentales. Le développement d'outils informatiques accessibles et faciles d'utilisation est crucial pour permettre aux biologistes d'effectuer leurs analyses post-expérimentales le plus efficacement possible. Dans cette suite d'idées, ce mémoire présente deux outils indépendants développés dans le cadre de projets biologiques bien précis. Bien que ces deux outils traitent des données de séquençages, un d'eux se rapporte au domaine de la génomique et l'autre au domaine de la génétique. Ce chapitre d'introduction présentera ainsi des éléments pertinents ainsi que notre hypothèse et les objectifs de travail de chacun des deux volets du mémoire.

1.1. Volet génomique

Le contexte de ce volet est l'étude de la terminaison de la transcription chez la levure à fission *Schizosaccharomyces pombe,* où un des mécanismes classiques de terminaison ne semble pas être actif. Quelques souches mutantes (Mut) pour des facteurs impliqués dans la terminaison seront utilisées, en particulier pour effectuer des expériences d'immunoprécipitation de chromatine couplée au séquençage (ChIP-Seq). La transcription d'un très grand nombre de gènes pouvant être affectée dans ces Mut par rapport à la souche sauvage (WT), il est nécessaire d'effectuer des expériences de

ChIP-Seq en ajoutant un contrôle externe (nommé ChIP-Seq-SI, pour *Spike-In*), permettant ainsi d'effectuer une normalisation corrigeant un potentiel biais causé par un effet global. Toutes ces notions seront introduites dans la présente section.

1.1.1. Les ARN polymérases

La transcription des génomes eucaryotes est effectuée par 3 complexes enzymatiques composés de plusieurs sous-unités, soit les ARN polymérases (ARNpol) I, II et III. Ils ont chacun une sensibilité différente pour la toxine alpha-amanitine, ce qui a permis d'identifier que chaque ARNpol transcrit différentes classes d'ARN cellulaires (Lindell *et al.*, 1970; Weinmann and Roeder, 1974; Zylber and Penman, 1971). Chez *Saccharomyces cerevisiae*, l'ARNpol I synthétise le précurseur de l'ARN ribosomal 25S, l'ARNpol II synthétise les ARN messagers (ARNm) et des ARN non-traduits et l'ARNpol III synthétise les ARN de transfert (ARNt) et l'ARN ribosomal 5S (ARNr). Il va sans dire que le fonctionnement de ces ARNpol dépend de plusieurs interactions avec des facteurs de transcription. Ces interactions sont en partie possibles grâce au domaine carboxyle terminal (CTD) des ARNpol (Vannini and Cramer, 2012).

Le CTD de l'ARNpol II est composé de 26 (levures) ou 52 (vertébrés) répétitions de l'heptapeptide Tyr-Ser-Pro-Thr-Ser-Pro-Ser (Corden, 1990). Ce domaine est essentiel pour la viabilité cellulaire, ce qui explique son haut taux de conservation chez les organismes eucaryotes. La phosphorylation de certains résidus du CTD permet de changer les interactions possibles de l'ARNpol II. Par exemple, la forme non-phosphorylée permet son recrutement par le complexe de préinitiation de la transcription. Les multitudes de modifications post-traductionnelles que le CTD subit lors d'un cycle de transcription sont la clef pour coordonner la succession du

recrutement des facteurs impliqués dans ce cycle (Hirose and Ohkuma, 2007; Laitem *et al.*, 2015).

1.1.2. La terminaison de la transcription

Le largage de l'ARN naissant et la dissociation de l'ARNpol II de l'ADN matrice sont des étapes fondamentales dans l'expression des gènes. En effet, un dérèglement de la terminaison de la transcription peut causer des interférences de transcription ou des collisions d'ARNpol en élongation, pouvant affecter l'expression des gènes avoisinants. Dans un génome ayant une composition génique hautement compacte, ces conséquences peuvent être encore plus prononcées (Jensen *et al.*, 2013; Shearwin *et al.*, 2005).

Chez les eucaryotes, la transcription et ses voies de terminaison ont une grande influence sur le sort des ARN transcrits. Bien que la terminaison de la transcription ait été étudiée dans plusieurs organismes, elle a été majoritairement étudiée chez *S. cerevisiae*. Deux principales voies de terminaison semblent être utilisées dépendamment du type de gène transcrit, soit la terminaison par le modèle nommé *Torpedo* ou par le complexe Nab1, Nrd1 et Sen1 (NNS) (Porrua and Libri, 2015).

1.1.2.1. Modèle *Torpedo*

Pour les gènes codants pour les ARNm, les données existantes supportent un mécanisme dépendant de la coupure du transcrit naissant par l'endonucléase Ysh1 qui fait partie du module nucléase du complexe de traitement de l'extrémité 3' (Casañal *et*

al., 2017) (Figure 1A). En effet, le recrutement cotranscriptionnel des facteurs de coupure et de polyadénylation de l'ARN au site de polyadénylation (site polyA) causent la coupure du pré-ARNm naissant, suivi par la polyadénylation de l'extrémité 3' ainsi formée sur le transcrit largué (Shi et Manley 2015). L'extrémité 5' se trouvant toujours liée à l'ARNpol II fournit une entrée vulnérable non coiffée pour l'exonucléase 5'-3' évolutivement conservée Rat1 (Dhp1 chez *S. Pombe*). La dégradation 5'-3' de l'ARN prend fin lorsque l'exonucléase rattrape l'ARNpol II et ainsi cause sa dissociation du brin d'ADN matrice (Baejen *et al.*, 2017; Fong *et al.*, 2015; Kim *et al.*, 2004; West *et al.*, 2004). Ce mécanisme est appelé la terminaison de la transcription médiée par *Torpedo*.

1.1.2.2. Complexe NNS

En plus des ARNm, l'ARNpol II synthétise une panoplie d'ARN non-codants (ARNnc) qui inclut les petits ARN nucléolaire (ARNsno), les petits ARN nucléaires (ARNsn), les longs ARN non-codants (ARNlnc) et des transcrits cryptiques instables (CUT). Chez *S. cerevisiae*, la terminaison de la transcription de ces gènes ne dépend pas de la machinerie de coupure/polyadénylation, mais utilise une voie sans coupure qui requière un complexe nommé NNS qui est constitué de protéines qui lient l'ARN (Nab3 et Nrd1) et d'une hélicase ADN/ARN (Sen1) (Porrua and Libri, 2015). Dans cette voie de terminaison, les protéines Nab1 et Nrd1 sont recrutées par la machinerie de transcription via le CTD grâce au domaine d'interaction du CTD (CID) de Nrd1 (Gudipati *et al.*, 2008; Vasiljeva *et al.*, 2008). Un motif spécifique enrichi en aval des gènes des ARNnc permet l'association de l'ARN naissant au complexe NNS créant ainsi un hybride ADN-ARN (Carroll *et al.*, 2004; Creamer *et al.*, 2011). Le recrutement de l'hélicase Sen1 au complexe d'élongation déstabilise l'hybride ADN-ARN, ce qui libère l'ARN naissant et la polymérase de l'ADN matrice (Porrua and Libri, 2013). Cette terminaison est couplée avec une addition d'une courte queue poly-adénosine (queue

polyA) par le complexe de polyadénylation TRAMP. Cette courte queue polyA est ciblée par le complexe de l'exosome d'exonuclease 3'-5' pour permettre à l'ARN d'être maturé (e.g. ARNsno) ou dégradé complètement (e.g. CUT) (Tudek *et al.*, 2014; Vasiljeva and Buratowski, 2006).

Par contre, chez *S. pombe* le complexe NNS ne semble pas actif. En effet, Nab3 et les homologues de Sen1 (Sen1 et Dbl8) du complexe NNS ne sont pas essentiels à la survie de la levure et l'homologue de Nrd1 (Seb1) n'interagit pas avec les autres membres du complexe (Lemay *et al.*, 2016). Toujours selon Lemay *et al.* (2016), Seb1 interagit avec les facteurs de terminaisons des ARNm, ce qui offre une piste intéressante pour ce qui est de la compréhension de la terminaison de transcription de l'ARNpol II aux gènes non-codants de la levure à fission.

1.1.2.3. CTD et terminaison

Parmi les modifications post-traductionnelles du CTD de l'ARNpol II, la phosphorylation de la sérine 2 (Ser2P) et de la sérine 5 (Ser5P) sont les plus étudiées et semblent être les plus abondantes chez *S. cerevisiae* (Suh *et al.*, 2016). Dans le cas des petits ARN non-codants, le recrutement du complexe NNS est influencé par Ser5P via le CID de Nrd1 (Gudipati *et al.*, 2008; Vasiljeva *et al.*, 2008). Au cours de l'élongation de la transcription des gènes codants, Ser5P est déphosphorylé et le taux de Ser2P augmente progressivement jusqu'au site de coupure/polyA où le taux de ser2P est à son plus haut (Kim *et al.*, 2010; Mayer *et al.*, 2010). Le facteur de terminaison Pcf11, reconnaît préférentiellement Ser2P via son domaine CID (Lunde *et al.*, 2010; Meinhart and Cramer, 2004), ce qui explique la forte localisation de ce facteur à l'extrémité 3' des gènes codants (Mayer *et al.*, 2012). Cependant, le recrutement de Pcf11 peut aussi être causé par une autre modification du CTD. Par exemple, il semble que chez *S. cerevisiae* Tyr1P peut empêcher le recrutement de Pcf11 par Ser2P *in vitro* (Mayer *et*

al., 2012). Schreieck *et al.* (2014) propose aussi que Tyr1P empêche le recrutement de Pcf11 dans la région codante du gène chez *S. cerevisiae* et que la chute du taux de Tyr1P légèrement en amont du site de coupure/polyA par Glc7 phosphatase permettrait à Pcf11 d'être recruté par Ser2P (Schreieck *et al.*, 2014).

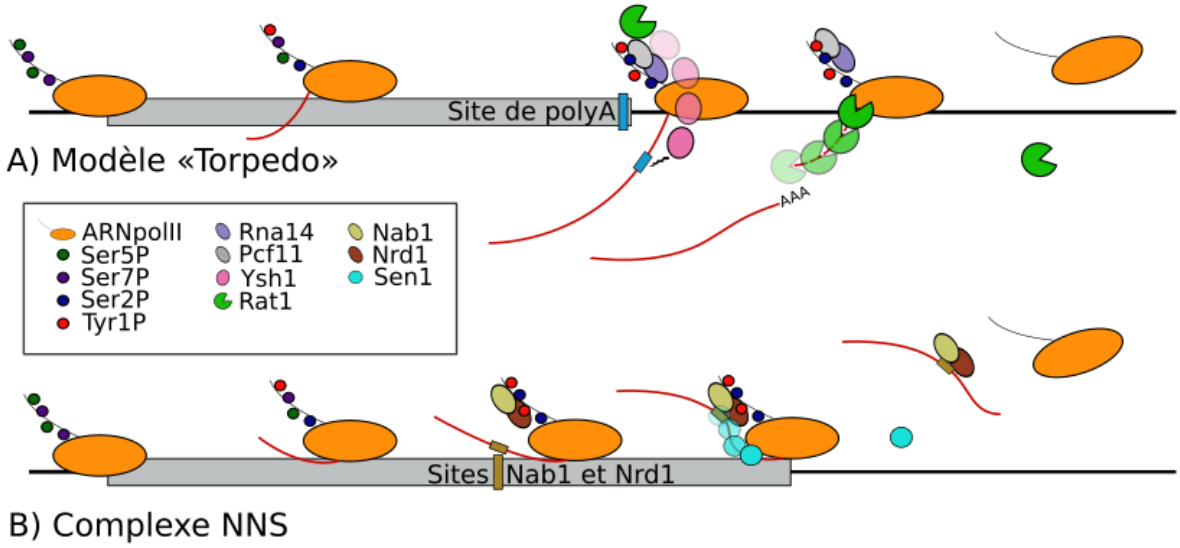


Figure 1. Schématisation des deux voies principales de terminaison de la transcription de l'ARNpol II chez *Saccharomyces cerevisiae*.
(A) Le modèle *Torpedo* est la voie de terminaison des ARNm, soit des gènes codants pour des protéines. La phosphorylation de la serine 2 du CTD permet le recrutement du complexe de facteurs de coupures (incluant Rna14, Pcf11, Ysh1, Rat1, etc.). Cette voie est dépendante de la coupure de l'ARN naissant au site polyA par l'endonucléase Ysh1 suivie par la dégradation de l'ARN toujours lié à l'ARNpol II par l'exonucléase 5'-3' Rat1. (B) Le complexe NNS est responsable de la terminaison de la transcription pour les gènes non-codants tel que les ARNsno, ARNsn, etc. Cette voie est indépendante d'une coupure, car la terminaison dépend de la perturbation de la stabilité du complexe ADN-ARN entre l'ADN matrice et l'ARN naissant par l'hélicase Sen1 recrutée par Nab1 et Nrd1 au CTD, puis relocalisée sur l'ARN naissant.

1.1.3. La méthode du ChIP-Seq-SI

1.1.3.1. Le ChIP-Seq standard

La technique de ChIP-Seq permet d'évaluer le taux d'occupation d'une protéine d'intérêt sur l'entièreté du génome en couplant la méthode de ChIP et les nouvelles technologies de séquençage (Barski *et al.*, 2007) (Figure 2). Les fragments d'ADN liant la protéine cible sont enrichis par immunoprécipitation et sont ensuite séquencés puis alignés sur le génome de référence (Figure 2). L'échantillon issu de ce protocole est généralement nommé IP (puisqu'il est le résultat d'une immunoprécipitation).

Généralement, un échantillon contrôle est généré ; les deux types les plus populaires sont l'*input* et le *mock*. L'*input* représente le matériel de départ prélevé entre la fragmentation et l'immunoprécipitation. Le *mock* représente les interactions non spécifiques et est généré par immunoprécipitation avec un anticorps ciblant un épitope non nucléaire tel que IgG. L'échantillon *mock* peut également être généré par immunoprécipitation avec le même anticorps que pour l'échantillon IP, mais dans une souche ne comportant pas l'épitope (e.g. l'échantillon IP est généré en utilisant un anticorps reconnaissant l'épitope HA ajouté à une protéine d'intérêt, l'échantillon *mock* serait généré avec l'anti-HA dans la souche parentale). En plus d'être utilisé dans certaines méthodes de normalisation telles que *Normalization of ChIP-Seq* (NCIS) (Liang and Keleş, 2012), l'utilisation d'un échantillon contrôle est fortement suggérée pour obtenir de meilleurs résultats avec les méthodes d'identification de régions statistiquement enrichies (*peak calling)* tel que *Model-based Analysis of ChIP-Seq* (MACS) (Zhang *et al.*, 2008). Le signal de l'échantillon contrôle est très souvent soustrait ou divisé du signal de l'IP.

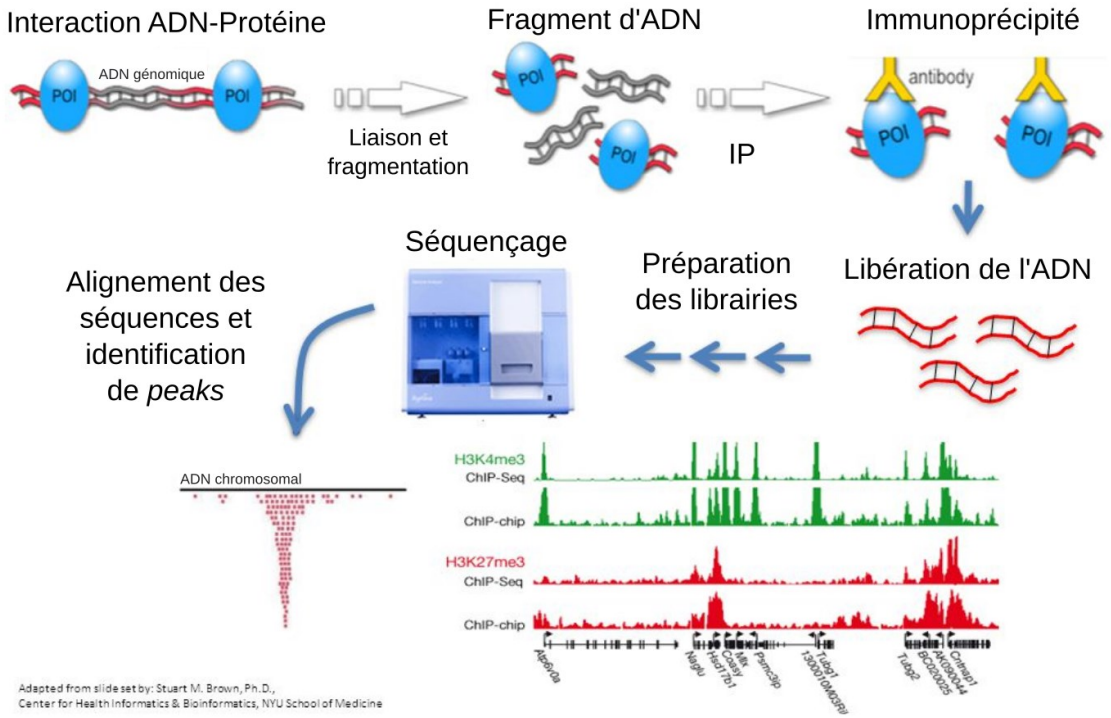


Figure 2. Schématisation des étapes d'une expérience de ChIP-Seq.
Cette technique permet d'évaluer les interactions directes et indirectes d'une protéine d'intérêt sur l'entièreté du génome étudié. Les étapes clés sont : stabilisation des interactions protéine-ADN par liaison covalente à l'aide de la formaldéhyde, fragmentation de l'ADN, enrichissement des cibles de la protéine d'intérêt par immunoprécipation, rupture des liaisons entre les fragments d'ADN et les protéines, préparation des banques d'ADN pour le séquençage, traitement bio-informatique et analyse des données (contrôle qualité, alignement, identification des sommets, etc.)

1.1.3.2. Fichiers et formats typiques utilisés dans l'analyse d'une expérience de ChIP-Seq

Assemblage : La séquence nucléotidique d'un organisme (génome de référence) est publiée le plus souvent sous forme incomplète et des ajustements sont ultérieurement effectués pour compléter ou corriger le génome de référence. Une nouvelle version du génome de référence est ainsi publiée chaque fois qu'une série d'altérations majeures

y sont apportées. Le nom de l'assemblage (e.g. hg19, sacCer3) fait référence à la version du génome de référence utilisée.

Annotation : Pour chaque assemblage, le fichier d'annotation contient les détails génomiques de chaque gène annoté. Le format BED est l'un des formats pouvant afficher l'annotation dans un navigateur de génome.

Format BED : Ce format permet de représenter une région génomique par ligne et sa flexibilité permet une caractérisation plus ou moins détaillée. Les trois colonnes minimales sont : chromosome, position chromosomique de début et position chromosomique de fin. Le nombre de colonnes varie de 3 à 12 (nommé BED3 à BED12). Le BED3 est le format utilisé pour représenter entre autres les régions statistiquement enrichies (*peaks*), alors que le BED12 est majoritairement utilisé pour représenter les annotations génomiques d'un assemblage. En effet, les autres colonnes d'un BED12 contiennent les détails d'affichage des régions sur le navigateur de région (le nom, le brin, les positions des régions non traduites, de la région codante et des exons).

Format bedGraph : Ce format permet de représenter et d'afficher sur un navigateur de génome une intensité de signal pour des régions de longueurs variables (e.g. ChIP-Seq). Ce format est composé d'un BED3 avec une quatrième colonne contenant la valeur, pouvant entre autres contenir la densité de lectures d'une expérience de séquençage.

Format WIG : Le format WIG (pour *wiggle*) est une alternative au format bedGraph prenant moins d'espace sur disque. Ceci est possible grâce à l'ajout d'une ligne

d'entête à chaque chromosome contenant optionnellement la longueur de chaque région, permettant ainsi de retirer la colonne de chromosome et la colonne de fin (qui est déduite par la valeur de début et la longueur) dans la configuration *variableStep* de ce format. Il est possible de réduire encore plus la taille de ce fichier en utilisant la configuration *fixedStep* où l'entête contient alors le nom du chromosome, la position de début et la longueur des pas, ne laissant ainsi que la colonne contenant les valeurs.

Format bigWig : Ce format est la version binaire indexée des formats bedGraph et WIG. En plus de prendre moins d'espace disque, l'index de ce format permet une accession très rapide à l'information. Ce dernier point est d'une très grande importance lors de l'affichage des données sur un navigateur de génome puisque seulement les données nécessaires sont transférées au navigateur. Il est ainsi conseillé d'utiliser le format bigWig plutôt que les formats bedGraph et WIG pour l'affichage sur les navigateurs de génome.

Les formats BED, bedGraph, WIG et bigWig ont tous été créés par l'équipe de UCSC [http://genome.ucsc.edu/FAQ/FAQformat.html].

1.1.3.3. Normalisation standard du signal

Une des utilisations courantes du ChIP-Seq est de comparer plusieurs expériences pour évaluer l'effet d'une variable (e.g. comparer les patrons de recrutement de facteurs d'activation dans différents tissus (Visel *et al.*, 2009), comparer les gènes transcrits par l'ARNpol II avant et après l'ajout d'une hormone (Belandia *et al.*, 2002)) sur le taux d'occupation d'une même protéine d'intérêt. Cependant, suite au séquençage, le nombre de lectures obtenues varie généralement (dans l'ordre des

millions), rendant la comparaison biaisée. La profondeur de séquençage est le principal facteur expliquant cette variabilité. Pour pallier cette limitation, une normalisation du signal est effectuée. Elle consiste à ramener chaque échantillon au même nombre de lectures en calculant un facteur de normalisation qui est ensuite appliqué à l'entièreté du signal (Bailey *et al.*, 2013). Le défaut de cette approche est l'assomption de la linéarité du biais causé par la profondeur de séquençage. Le développement d'outils de normalisation non linéaire est de plus en plus populaire. (Liang and Keleş, 2012; Taslim *et al.*, 2009).

1.1.3.4. Détection d'un effet global

Tel que mentionné, la normalisation standard peut dans certains cas dissimuler des conclusions biologiques importantes, comme un effet global. En effet, une condition expérimentale créant une diminution ou une augmentation de l'enrichissement d'une protéine à travers l'entièreté du génome, devrait faire varier l'intensité du signal dans la même direction. Cependant, la normalisation par le nombre total de lectures alignées a pour effet de ramener le signal d'une telle expérience au même niveau que l'expérience contrôle (Figure 3A) (Bonhoure *et al.*, 2014; Orlando *et al.*, 2014). Il est attendu que la mutation d'un facteur général de transcription aura un effet global sur l'enrichissement de l'ARNpol II. Il est à noter que de tels effets globaux ont déjà été identifiés et pris en compte lors de l'analyse par l'utilisation d'une référence externe dans le cas d'études transcriptomiques sur puce d'ADN (van de Peppel *et al.*, 2003) et par séquençage (RNA-seq) (Lovén *et al.*, 2012), ainsi que dans le cas d'étude de ChIP-chip (Jeronimo *et al.*, 2015) où les fragments d'ADN sont hybridés sur puce plutôt que séquencés.

1.1.3.5. Normalisation par *Spike in* (SI)

Une variante de la technique de ChIP-Seq utilisant une référence externe a ainsi été récemment proposée pour pallier le biais causé par la normalisation standard (Bonhoure *et al.*, 2014; Orlando *et al.*, 2014). Cette méthode nommée *Spike-in* (SI), est une technique expérimentale apportant une alternative de normalisation. Le SI consiste à mettre une proportion fixe de chromatine d'une espèce exogène de référence dans les tubes de chacune de nos expériences. Il sera donc possible de normaliser avec le nombre de lectures alignées sur le génome de l'espèce exogène, puisque la chromatine exogène sera dans le même état dans toutes les conditions expérimentales (Figure 3B). Le choix de l'espèce exogène comporte des défis en soi. En effet, les deux espèces doivent être proches phylogénétiquement pour permettre à un même anticorps de reconnaître l'épitope de la protéine cible chez les deux espèces lors de l'immunoprécipitation. Par contre, cette distance évolutive ne doit pas être trop courte, car on veut empêcher l'alignement d'une lecture provenant de chromatine exogène sur le génome de l'espèce étudié et vice-versa (Bonhoure *et al.*, 2014; Orlando *et al.*, 2014).

Les publications de Orlando *et al.* (2014) et Bonhoure *et al.* (2014) proposent chacune une méthode de normalisation semblable, mais non identique. Dans les deux cas, la normalisation consiste bien entendu à exprimer le signal de l'expérience de façon relative au signal du génome exogène présent. Les différences majeures entre les deux approches sont les suivantes :

- Orlando prend un génome chimérique alors que Bonhoure aligne sur les deux génomes séparément et supprime les lectures qui alignent sur les deux génomes.

- Orlando prend en compte le nombre de lectures alignées sur la totalité du génome exogène pour effectuer la normalisation alors que Bonhoure prend en compte seulement les lectures dans les régions enrichies du génome exogène.
- Orlando normalise le signal d'*input* de la même façon que le signal expérimental, puis les soustrait (donc normalise par SI les *inputs* et les expériences, puis soustrait l'*input*), alors que Bonhoure soustrait l'*input* aux IP (suite à une normalisation standard) avant de normaliser par le SI (à noter que les résultats de la soustraction de l'*input* sont ramenés à zéro s'ils sont négatifs).

1.1.5. Hypothèse et objectif général du volet génomique

Les évidences pointant que le complexe NNS n'est pas fonctionnel chez *S. pombe* laissent suggérer que la terminaison de la transcription des ARN non-codants est différente chez les deux levures. Nous croyons que la terminaison utilisant le modèle *Torpedo* pourrait être utilisée pour les ARN non-codants chez la levure à fission étant donné que sa seule homologue essentielle du complexe NNS est Seb1 et qui semble interagir avec les facteurs de terminaison impliquée dans le modèle *Torpedo*. Pour investiguer cette hypothèse, des Mut dans le complexe de coupure/polyadénylation seront utilisés. Cependant, des résultats préliminaires ont suggéré une diminution globale de l'enrichissement de l'ARNpol II sur tout le génome chez ses Mut. Mon objectif principal dans ce projet est ainsi de mettre au point et utiliser un outil de normalisation nommé SpkNorm afin d'analyser des données de ChIP-Seq-SI.

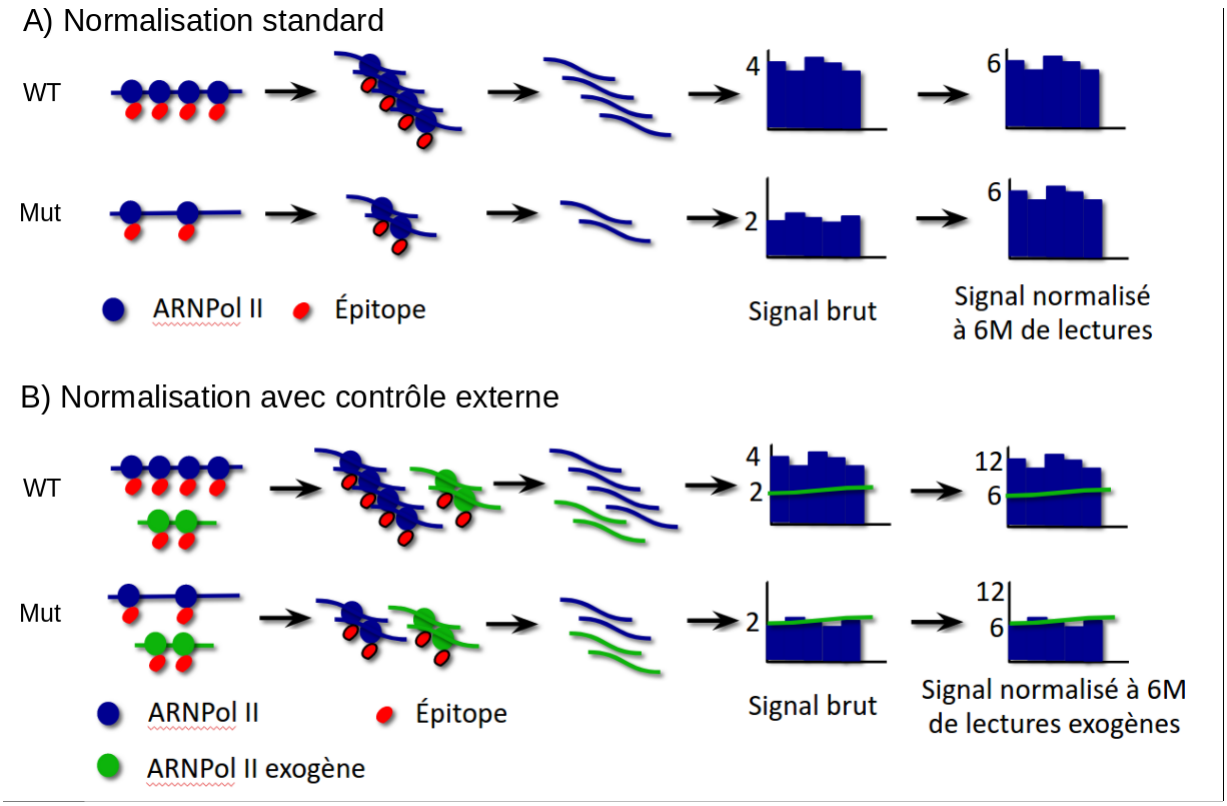


Figure 3. Schématisation de la limitation de la normalisation standard lors d'effet global dans une souche mutante et application de la normalisation avec SI.
(A) Normalisation standard d'une souche sauvage (WT) et souche mutante (Mut) où l'ARNpol II est globalement deux fois moins présente sur l'ADN. Suite au ChIP-Seq, le signal brut du WT est deux fois plus élevé que celui du Mut. La normalisation standard ramène les expériences à un nombre égal de lectures pour que les résultats soient comparables (ici six millions de lectures). Par contre, une telle normalisation vient complètement camoufler l'effet de diminution biologique. (B) Normalisation avec chromatine exogène (en noir) pour les mêmes souches qu'en A. Suite au ChIP-Seq, les lectures qui alignent sur le génome de l'organisme exogène peuvent servir de référence. La normalisation utilise donc ces lectures pour ramener les expériences à un nombre égal de lectures alignées sur le génome exogène (ici six millions). Une telle normalisation permet ainsi de détecter un effet biologique global.

1.2. Volet génétique

Le contexte de ce volet est une potentielle application de médecine personnalisée visant à déterminer les prédispositions d'un fœtus de développer des maladies génétiques récessives rares et de fournir aux futurs parents les informations nécessaires pour prendre une décision éclairée quant aux potentiels problèmes de l'enfant à naître. Outre les questions éthiques, une telle approche aura un impact économique important. Pour évaluer cet impact, le nombre de couples nécessaires pour détecter suffisamment de couples à risque est une réponse biologique cruciale. L'utilisation de milliers de fichiers *variant call format* (VCF) ou de fréquences de variants populationnels sont deux ressources potentielles de réponses. Toutes ces notions seront introduites dans cette présente section.

1.2.1. Variations génétiques

Le développement des techniques de séquençage a permis à plusieurs projets génétiques de voir le jour, tel que *1000 genomes project* qui a séquencé le l'exome de 2504 individus (Gibbs *et al.*, 2015). Ils estiment que la séquence d'un génome typique diffère au génome de référence de 4.1 à 5 millions sites. Bien que plus de 99.9 % de ces variants sont des polymorphismes d'un seul nucléotide (SNP) et des courtes insertions ou délétions, les variants structuraux couvrent beaucoup plus de bases, soit environ 20 millions de bases (Gibbs *et al.*, 2015).

De plus, seulement 1 à 4 % des variants d'un individu se retrouvent chez moins 0.5 % de la population (Gibbs *et al.*, 2015). La fréquence allélique de ces 88 millions de variants identifiés par le *1000 genomes project* sont utilisées comme banque de

données (valeur CAF) dans l'outil d'annotation de variant tel que dbSNP (Sherry *et al.*, 2001). De plus, le *Exome Aggrégate Consortium* (ExAC), maintenant *Genome Aggregation database* (gnomAD) détient une banque de données des fréquences alléliques plus complète que dbSNP étant donné l'incorporation de plusieurs autres projets de séquençages, montant ainsi la quantité de données utilisées à 23,136 exomes et 15,496 génomes (Lek *et al.*, 2016).

1.2.2. Maladies génétiques récessives rares

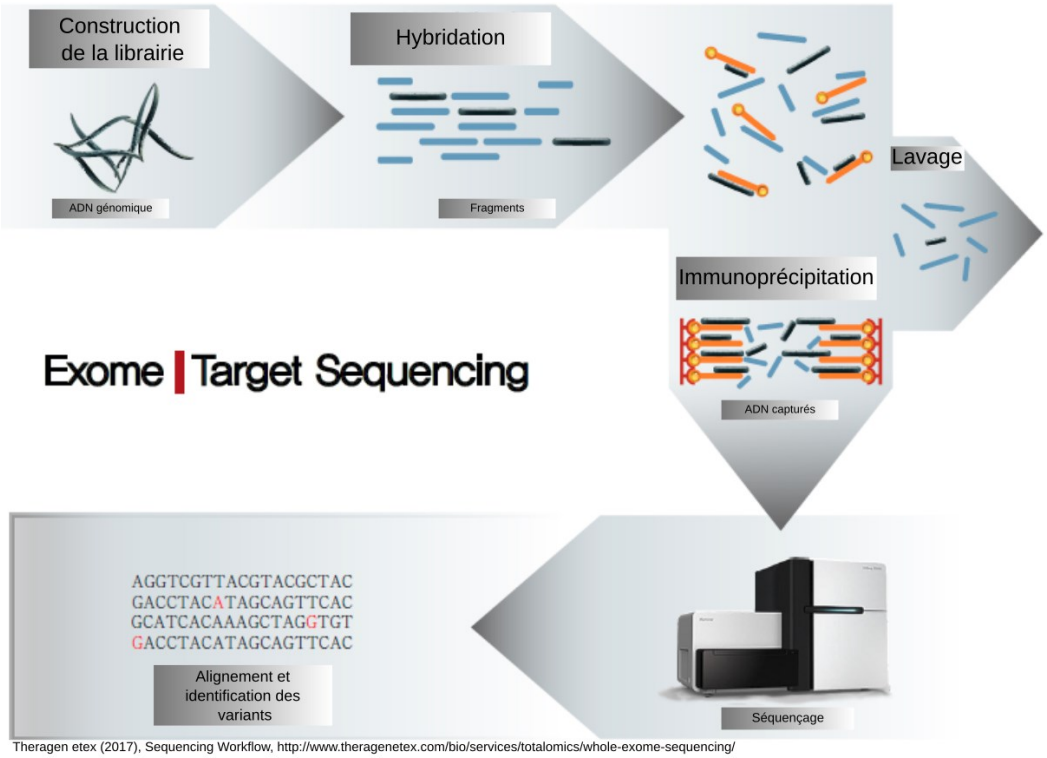
Bien que la majorité des variants ne cause aucun impact, il est possible qu'un variant (ou un ensemble de variants) soit responsable d'un trait phénotypique variable. Par exemple, chez l'humain, la perte de la capacité de goûter l'amertume à certains aliments est causée par trois variants localisés dans le gène TAS2R38 (Deshaware and Singhal, 2017). En effet, l'un de ces trois variants est suffisant pour rendre inactive la protéine G transmembranaire qui régit le goût du glucosinolate. Bien que les variants soient cruciaux pour une bonne diversité génétique, ils sont aussi responsables pour les maladies génétiques, on dira d'un tel variant qu'il est pathogénique. S'il est sur un chromosome autosomal, on dira que la maladie est autosomale ; dans la même logique, on parle d'une maladie liée au chromosome X ou Y lorsque le variant est sur l'un des chromosomes sexuels. Dans le cas des maladies autosomales récessives (e.g. syndrome de Fowler (Lalonde *et al.*, 2010)), un individu hétérozygote aura un phénotype normal; ces individus sont nommés porteurs sains. Cependant, un individu homozygote pour l'allèle pathogénique développera la pathologie. Dans le cas des maladies autosomales dominantes (e.g. syndrome Kabuki (Ng *et al.*, 2010)), la présence du variant pathogénique est la garantie d'une pathologie. En effet, un individu hétérozygote ou homozygote développera la pathologie. La banque de données ClinVar répertorie les découvertes liant un ou plusieurs variants à un phénotype et ainsi contient l'information sur la pathogénicité de plusieurs variants (Landrum *et al.*, 2018).

ClinVar étant limité aux variants connus, l'utilisation d'outils permettant de prédire l'impact d'un variant quelconque sur son ou ses gène(s) associé(s) permet d'utiliser des informations sur les variants absents de ClinVar. En effet, il est possible de prédire l'impact d'un variant en identifiant quel genre de mutation que le variant engendre au niveau de la séquence protéique. Par exemple, l'apparition prématurée d'un codon d'arrêt de la traduction permet de prédire un impact élevé du variant sur la fonction de la protéine. Un des outils les plus populaires pour ce faire est snpEff (Cingolani *et al.*, 2012). Dans le cadre de ce mémoire, seulement les variants autosomaux récessifs pathogéniques rares (donc ayant une fréquence allélique < 1 % selon ExAC) seront considérés.

1.2.3. Séquençage complet de l'exome

Au cours des dernières années, l'étude des maladies génétiques est devenue plus populaire à cause de l'apparition de méthodes plus puissantes pour l'identification de variants. En effet, grâce à la grande avancée des techniques de séquençage, le séquençage complet de l'exome (WES) est devenu une expérience raisonnable et suffisamment abordable pour identifier facilement et rapidement les variants des exons de plusieurs individus (Ku *et al.*, 2011). Cette technique consiste à enrichir les régions exoniques connues d'un échantillon d'ADN par hybridation avec des sondes complémentaires à tous les exons. Cet ADN est ensuite récupéré puis préparé pour le séquençage (voir Figure 4). À la suite du séquençage et de l'analyse bio-informatique, il est possible d'explorer la totalité des variants de bonne qualité contenu dans les exons de l'individu et présentés dans un fichier au format (VCF, *Variant Call Format*). Ce type de fichier est un format texte possédant un entête contenant des informations sur les traitements appliqués (filtres, version des annotations, etc.), suivi d'une ligne par variant, contenant toutes ses informations de bases (chromosome, position, allèle référence et alternative) ainsi que toutes annotations ajoutées par les outils utilisés

(fréquences alléliques, gènes associés, impacts prédits, etc.) [http://www.internationalgenome.org/wiki/Analysis/vcf4.0/]. De façon générale on retrouve ~4000 variants dans les exons d'un individu (Trakadis *et al.*, 2014), dont ~500 variants rares de bonne qualité. L'utilisation d'un outil tel que PhenoVar permet ensuite d'utiliser les différentes annotations ci-haut et les phénotypes de l'individu pour prioriser les variants jusqu'à identifier ~15 variants potentiellement causaux (Thuriot *et al.*, 2018).



Theragen etex (2017), Sequencing Workflow, http://www.theragenetex.com/bio/services/totalomics/whole-exome-sequencing/

Figure 4. Schématisation des étapes d'une expérience de WES.
Les étapes schématisées représentent les étapes clés simplifiées. Ces étapes sont : fragmentation de l'ADN et préparation des librairies, hybridation des fragments à des sondes ADN étiquetées (peut aussi être fait dans des microplaques où les sondes sont liées à la plaque), immunoprécipitation grâce à l'étiquette et lavage, séquençage et traitement bio-informatique (alignement, identification des variants).

1.2.4. Hypothèse et objectif général du volet génétique

L'augmentation de l'accessibilité des techniques de séquençage n'ouvre pas seulement des portes dans le domaine de la recherche. En effet, le WES permet souvent d'expliquer dans un cadre clinique la cause d'une maladie génétique et éventuellement la prévenir et/ou la traiter. Les biologistes et les médecins commencent à développer des moyens d'utiliser ces nouvelles informations pour mieux traiter les patients. Le développement de ces nouvelles approches est nommé la médecine personnalisée. Une de ces applications est de déterminer les prédispositions d'un fœtus de développer des maladies génétiques récessives rares et de fournir aux futurs parents les informations nécessaires pour prendre une décision éclairée quant aux potentiels problèmes de l'enfant à naître. Il est bien entendu que de telles approches apportent beaucoup de questions éthiques, et peuvent également avoir un impact économique. Le chapitre 3 de ce mémoire s'affaire à donner des pistes sur le meilleur moyen d'estimer combien de couples il faudrait tester pour détecter un nombre suffisant de couples à risque. L'hypothèse principale est que d'utiliser des vraies données de WES donnera des résultats plus valables que l'utilisation des fréquences de variants populationnelles. Cette information pourra ensuite être utilisée pour mettre en relation les coûts du WES avec les coûts du traitement de ces maladies pour la société.

1.3. Objectifs spécifiques

L'objectif global de ma maitrise est de développer et d'utiliser des outils bio-informatiques dans le cadre de deux projets utilisant des données de séquençage dans des contextes de génomique et de génétique.

Pour le projet portant sur la génomique, la nécessité d'utiliser de la chromatine exogène pour effectuer une normalisation appropriée et le manque d'outils bio-informatiques permettant d'effectuer une telle normalisation sont les motivations principales pour développer l'outil SpkNorm, permettant de normaliser des données de ChIP-Seq-SI. Le chapitre 2 décrit les différentes étapes pour y parvenir.

Concernant le projet portant sur la génétique, il est d'un grand intérêt de comparer l'évaluation du nombre de couples à risque d'engendrer une progéniture atteinte d'une maladie génétique récessive rare à partir de listes de variants d'individus réels, par rapport à utiliser les fréquences populationnelles. Le chapitre 3 décrit ainsi le développement de l'outil modulable et paramétré MockScreen permettant de comparer ces résultats, ainsi que d'évaluer plus précisément l'impact d'un séquençage systématique des parents en vue de les informer sur la prédisposition génétique de leurs futurs enfants.

CHAPITRE 2
SPKNORM : UN OUTIL DE NORMALISATION DE CHIP-SEQ-SI ET SON APPLICATION LORS D'UNE ÉTUDE SUR LA TERMINAISON DE LA TRANSCRIPTION CHEZ *S. POMBE*

Dans le cadre de ce chapitre, des données génomiques générées en collaboration avec le laboratoire du Dr. François Bachand utilisant entre autres la méthode de ChIP-Seq-SI seront analysées, dans le but de mieux comprendre le processus de terminaison de la transcription chez *S. pombe*. La section 2.1 décrit les aspects plus techniques du développement de l'outil SpkNorm utilisé dans ce projet. La section 2.2 consiste quant à elle au manuscrit intitulé *Common mechanism of transcription termination at coding and noncoding RNA genes in fission yeast* soumis au journal *Nature communication*.

2.1. Mise en place et développement

Cette section explore et met sur la table les décisions qui ont été prises lors de son développement. Le cœur de l'outil SpkNorm a été implémenté en python3.4.0 et est intégré dans un script écrit en bash le liant avec les autres outils.

2.1.1. Choix de l'organisme externe

Pour effectuer une expérience de ChIP-Seq-SI, l'organisme externe doit être suffisamment éloigné afin que ses lectures ne puissent pas (ou peu) s'aligner sur le génome étudié et vice-versa. La présente section porte sur l'évaluation de *S. cerevisiae*

comme source de chromatine exogène pour l'application du SI dans des expériences de ChIP-Seq chez *S. pombe*.

Pour évaluer si les deux levures sont assez éloignées pour éviter des alignements sur le mauvais génome, des lectures synthétiques ont été générées à partir de la séquence du génome de référence de chaque levure. Les assemblages ASM294v2.29 et sacCer3 ont été utilisés pour générer des lectures de 100 paires de bases (pb), correspondant à la longueur attendue de la plupart des lectures générées dans ce projet. Les lectures étant décalées de 1 pb, à chaque position du génome il est attendu de retrouver 100 lectures alignées (sauf aux extrémités bien entendu). Deux des outils d'alignement les plus populaires ont été testés : BWA version 0.6.2 avec l'option « mem » et les paramètres par défaut (Li and Durbin, 2010), et Bowtie2 version 2.2.5 avec l'option « sensitive » (donc paramètres -D 15 -R 2 -N 0 -L 22 -i S 1,1.15) (Langmead and Salzberg, 2012). Pour calculer la quantité de lectures à chaque position du génome, nous avons ensuite utilisé la fonction genomeCoverage de l'outil BEDTools (Quinlan and Hall, 2010). Il a ainsi été possible de calculer la proportion de chaque génome contenant une certaine densité de signal (Tableau 1).

Tableau 1. Caractérisation de l'alignement des lectures synthétiques sur le génome de *S. pombe*.

Origine des lectures	Outil	Nombre de pb de *S. pombe* avec un signal				% aligné
		0	<100	100	>100	
S. pombe	BWA	391 <0,01%	24 274 0,20%	12 142 811 97,84%	242 903 1,96%	99,99%
	Bowtie 2	305 <0,01%	26 384 0,21%	12 112 642 97,72%	256 048 2,06%	99,99%
S. cerevisiae	BWA	12 600 811 99,76%	27 896 0,22%	103 <0,01%	2569 0,02%	0,27%
	Bowtie 2	12 617 904 99,90%	12 638 0,10%	68 <0,01%	769 <0,01%	0,40%

Tel qu'attendu, les lectures de *S. pombe* s'alignent majoritairement au bon endroit sur le génome de *S. pombe* (>95% des positions du génome possèdent exactement un signal de 100), alors que les aligneurs n'arrivent pas à placer de lecture sur <0.01% du génome ; ces régions correspondent essentiellement aux quatre régions de 99 pb composées de N (majoritairement dans les régions centromériques), puisque le génome n'a pas encore été parfaitement séquencé. Ainsi, seulement ~250 kilobase (kb) du génome de *S. pombe* possèdent un signal >100, vraisemblablement causé par des séquences répétées dans le génome. À l'inverse, une très faible proportion des lectures de *S. cerevisiae* peuvent être alignées sur *S. pombe*, >99.7% des positions de *S. pombe* possédant ainsi exactement un signal de 0. De plus, il est fort probable qu'une fraction importante des lectures de *S. cerevisiae* qui ont réussi à s'aligner sur *S. pombe* pourrait être mieux alignée sur son génome d'origine.

Pour tester cette hypothèse, nous avons ainsi aligné les 24 786 184 lectures synthétiques provenant des deux génomes sur un génome chimérique composé des deux levures (les chromosomes de *S. cerevisiae* ayant un le préfix « chr » sont facilement différenciables). Tel qu'attendu, la quasi-totalité des lectures de *S. cerevisiae* capable d'être alignées sur le génome de *S. pombe* sont préférentiellement alignées sur leur propre génome, <350 lectures étant alignées sur le mauvais génome (Tableau 2). À la lueur de ces résultats, les lectures générées à partir du génome de *S. cerevisiae* semblent très peu interférer avec le signal de *S. pombe* et vice-versa, faisant de ces deux levures d'excellents candidats pour une utilisation dans un contexte de SI. L'utilisation d'un génome chimérique aidant grandement à réduire l'interférence, c'est ce que nous avons choisi pour ce projet.

Tableau 2. Nombre de lectures synthétiques alignées sur le génome chimérique des deux levures.

*Lectures alignées sur :	Outils	Lectures provenant de :	
		S. pombe	*S. cerevisiae*
S. pombe	BWA	12 630 385	167
	Bowtie 2	12 629 936	166
S. cerevisiae	BWA	158	12 155 238
	Bowtie 2	167	12 155 239

* Lectures non-alignées : 236 pour BWA et 676 pour Bowtie2.

Les résultats sont très similaires entre les deux aligneurs, et encore plus en utilisant le génome chimérique. La différence majeure entre les deux aligneurs semble se retrouver dans l'alignement de lectures sur les frontières des quatre régions de 99 N du génome de *S. pombe*. Pour comprendre un peu mieux ce phénomène, nous avons ainsi utilisé une section de 1000 bp du génome de *S. pombe* possédant un signal parfait de 100 lectures, où nous avons remplacé les 200 bp du centre par des N, puis aligné toutes les séquences avec les deux outils et calculé le signal tel que précédemment (Figure 5). Tel que l'illustre la Figure 5, BWA permet de couper (*soft clip*) les lectures jusqu'à une longueur de 30 pb afin de les aligner, tandis que Bowtie2 ne coupe pas les lectures et permet un maximum de 15 pb consécutives non-complémentaires. Considérant cette légère meilleure sensibilité de BWA et considérant les résultats semblables d'alignement présentés précédemment, BWA est l'aligneur utilisé dans SpkNorm.

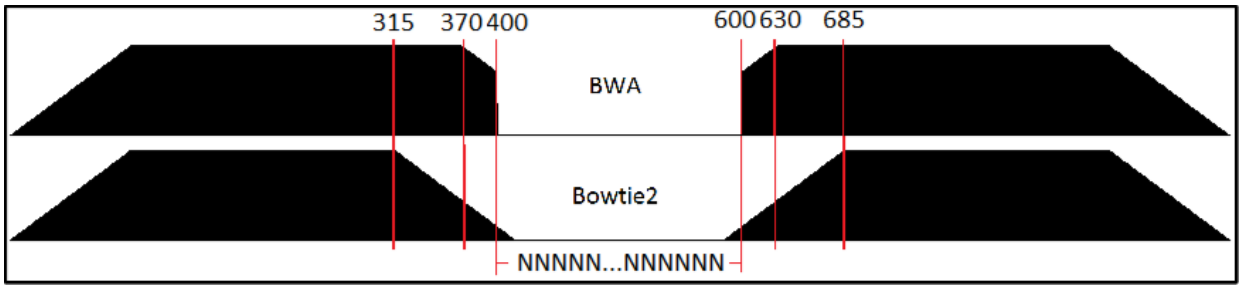


Figure 5. Distribution des lectures alignées par BWA et Bowtie2 sur une région de 1 kb contenant 200 pb de N en son centre.
Des lectures synthétiques de 100 pb commençant à toutes les positions de cette construction ont été générées. Les plateaux de ces signaux correspondent donc à des alignements parfaits. Les lignes rouges correspondent aux positions où l'alignement des lectures est affecté par la présence des N.

2.1.2. Détermination de la quantité optimale de chromatine exogène

Afin de déterminer la concentration de chromatine exogène à utiliser pour nos expériences de ChIP-Seq-SI, une analyse de saturation a été effectuée. À partir d'un ensemble de données provenant d'une expérience de ChIP-Seq ciblant l'ARNpol II chez *S. pombe* du WT sans SI, plusieurs ensembles de lectures synthétiques ont été générés en enlevant aléatoirement une certaine proportion des lectures (de 5% à 90% avec intervalles de 5%). Ces lectures ont été alignées avec BWA-mem, le signal moyen sur chaque gène a été calculé à l'aide de VAP version 1.1.0 (Brunelle *et al.*, 2015; Coulombe *et al.*, 2014), puis une normalisation a été effectuée pour tenir compte du nombre total de lectures alignées dans chaque ensemble de données afin de calculer l'équivalent d'une valeur RPKM d'enrichissement d'ARN pol II par gène.

Les gènes ont ensuite été regroupés en quartile en fonction de leur valeur RPKM dans l'ensemble de données sans réduction (contenant 100% des lectures). Cette

séparation permet de voir l'impact de la dilution synthétique sur quatre groupes de gènes, où le quartile de gènes ayant les plus faibles valeurs RPKM est celui qui sera le plus affecté par le retrait des lectures. L'erreur relative a ensuite été calculée pour chaque gène en faisant la soustraction de la valeur RPKM obtenue dans l'ensemble de données non réduit par la valeur observée dans l'ensemble réduit.
La distribution des erreurs relatives des gènes du dernier quartile de *S. pombe* montre un faible impact des premières dilutions dans les ensembles de données réduits pour une expérience sans SI, supportant que les données soient presque saturées (Figure 6A). En effet, retirer jusqu'à ~40-50% des lectures ne semble pas affecter significativement la distribution des valeurs RPKM (4,2% d'erreur relative à 50%). Tel qu'attendu, le graphique des gènes de *S. cerevisiae* sans SI (donc où la majorité des gènes n'a aucun signal) est un cas extrême d'expérience non saturée (Figure 6B).

Trois expériences de ChIP-Seq-SI ont été effectuées avec des concentrations de 2,5%, 5% et 10% de SI. L'objectif est d'utiliser la plus haute concentration possible de SI n'altérant pas la saturation du signal sur *S. pombe*, tout en permettant d'utiliser les lectures de *S. cerevisiae*. L'analyse de saturation de l'ensemble de données provenant de l'expérience avec 10% de SI a permis de confirmer que sa saturation est semblable à l'ensemble de données sans SI sur *S. pombe* (Figure 6A et C) (5,4% d'erreur relative à 50% de réduction, soit 1,2% de plus qu'en absence de SI); les données à 2.5% et 5% étaient bien entendu un peu plus saturées pour *S. pombe*. Cependant, le nombre de lectures sur les gènes du dernier quartile de *S. cerevisiae* n'est clairement pas saturé, même à 10% de SI (Figure 6D) (10,8% d'erreur relative à 50% de réduction). Nous avons néanmoins opté pour ne pas diluer plus la chromatine de *S. pombe*, en sachant qu'il sera possible d'augmenter la saturation de *S. cerevisiae* en combinant les lectures des différentes expériences (puisqu'elles utiliseront toutes le même lot de chromatine exogène).

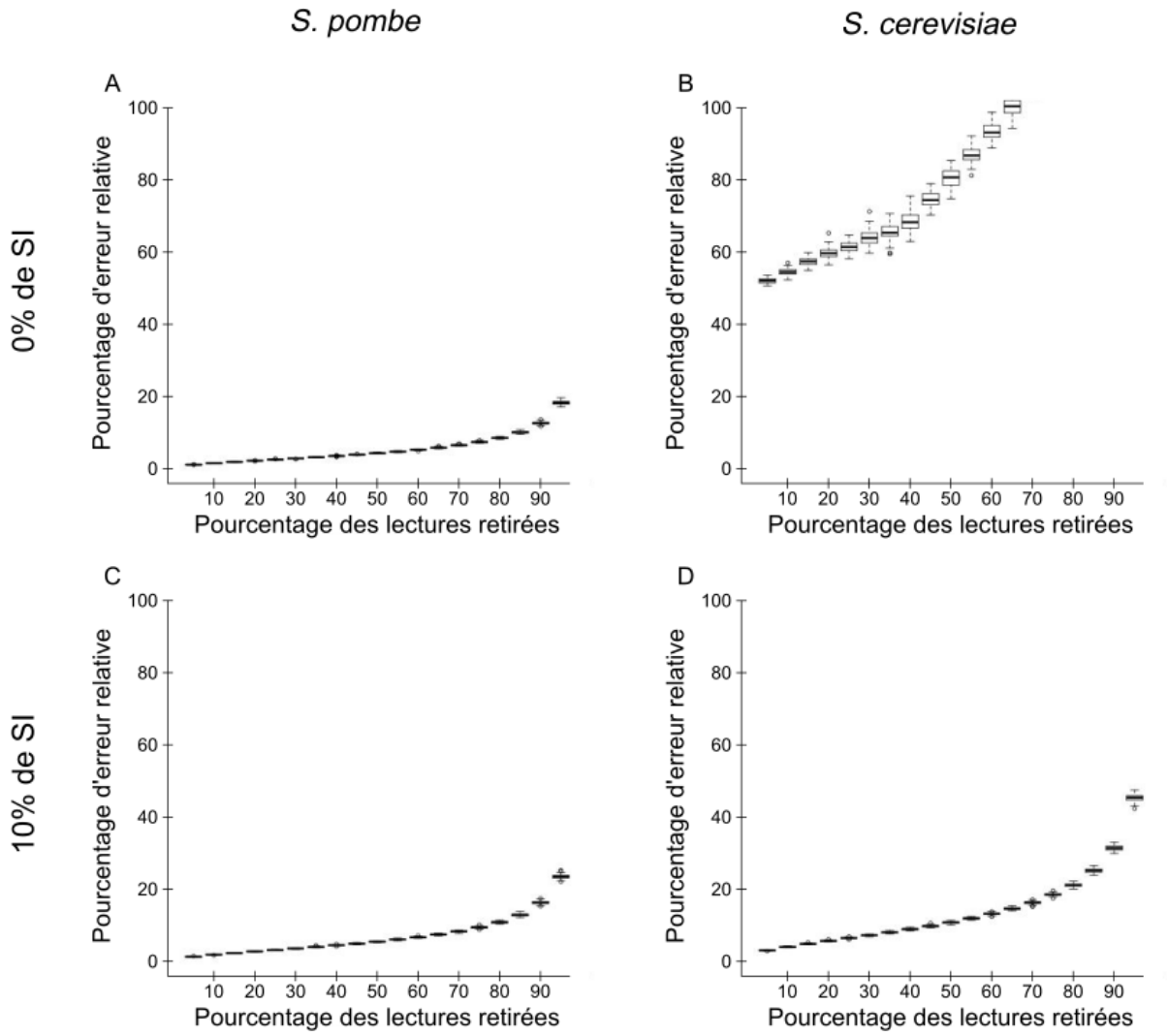


Figure 6. Analyse de saturation.

Analyse de saturation des gènes du quartile avec le plus faible signal d'ARNpol II chez *S. pombe* (A et C) ou *S. cerevisiae* (B et D) lors d'une expérience avec 0% (A-B) ou 10% (C-D) de chromatine de *S. cerevisiae* comme SI. Les graphiques représentent la distribution des pourcentages d'erreur relative des RPKM de chaque gène des différentes réductions par rapport à aucune réduction (donc par rapport aux RPKM calculés sans retirer de lecture). À noter que pour l'analyse sans SI, 87% des gènes de *S. cerevisiae* (B) ont été écartés de l'analyse puisque leur RPKM est a 0 (car très très peu de lectures alignées sans SI), alors qu'aucun gène n'a été exclu pour les autres analyses. Ce graphique a été généré à l'aide du logiciel R.

2.1.3. Préparation des fichiers de densité pour la normalisation

À la suite de l'alignement des lectures, l'étape suivante est généralement de créer un fichier de densité qui rapporte le nombre de lectures à chaque position du génome. Un des outils les plus populaires pour ce faire est la fonction genomeCovrage de BEDTools (Quinlan and Hall, 2010) qui sort la densité au format BedGraph, où chaque région a une taille variable (Figure 7). Cependant, pour faciliter la normalisation des fichiers, ce signal doit être fragmenté en région d'une taille définie (généralement 10 pb) souvent représenté par le format de fichier WIG (*fixed step*). La fonction multicov de BEDTools version 2.17.0, la fonction map de BEDTools ainsi que le script python « bedgraph_to_wig.py » [https://gist.github.com/svigneau/8846527] ont été testés. La fonction Multicov calcule simplement le nombre de lectures d'un fichier BAM chevauchant chaque région d'un fichier BED d'intérêt, apportant une certaine imprécision (e.g. région 51-61 où la majorité de la région n'a aucun signal mais obtient une valeur de deux lectures). La fonction Map permet d'effectuer une opération sommaire (e.g. max, min, moyenne, médiane) entre les régions d'un fichier de signal bedGraph et d'un fichier BED d'intérêt, de sorte que la moyenne des lectures ignore la longueur relative de chaque sous-région participante (e.g. région 41-51 où la moyenne est obtenue par le nombre de lectures chevauchant chacune des trois sous-régions donc (1 + 3 + 2) / 3 régions = 2 lectures). Le script bedgraph_to_wig calcule quant à lui la moyenne pondérée du signal d'un fichier bedGraph sur des régions d'une longueur donnée couvrant tout le génome (e.g. région 41-51 où le nombre de bases que chaque signal couvre est pris en compte (1*3pb + 3*2pb + 2*5pb) / 10pb = 1.9). SpkNorm utilise bedgraph_to_wig.py puisque la moyenne pondérée est la méthode la plus convenable et la plus rapide.

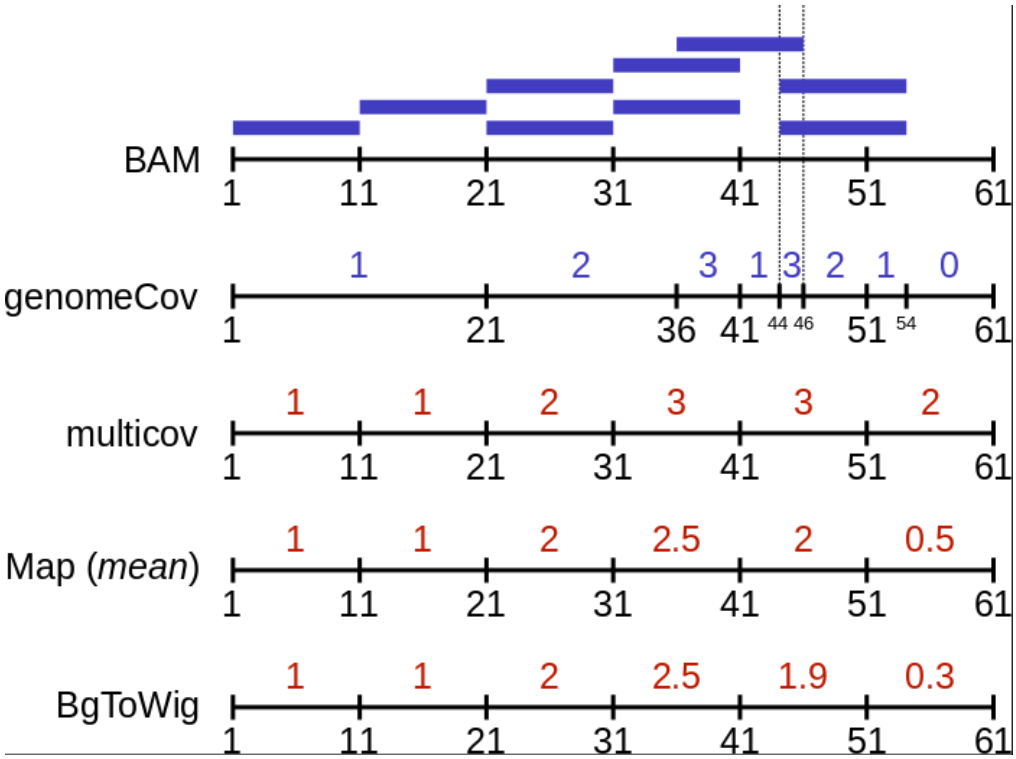


Figure 7. Comparaison des méthodes de fragmentation uniforme du signal. La première ligne contient la position de neuf lectures (représentées par les barres bleues) sur une portion de génome. Les lignes suivantes correspondent à la densité de lectures à la sortie de quatre différentes méthodes.

2.1.4. Normalisation par SI

Tel que mentionné précédemment, les approches d'Orlando et de Bonhoure diffèrent quant à la façon d'utiliser l'*input*. En effet, Bonhoure propose de soustraire le signal de l'*input* avant la normalisation par SI, ce qui implique de calculer la somme du signal seulement dans les régions supérieures à l'*input*, donc indirectement seulement dans les régions enrichies. Pour cette raison, deux variantes ont été testées : normalisation par la somme du signal SI avant la soustraction de l'*input* (l'*input* est normalisé de la même façon avant d'être soustrait, Orlando) et normalisation par la somme du signal

SI après cette soustraction (Bonhoure). Dans l'approche de Bonhoure, pour éviter que les régions où le signal de l'*input* est plus élevé que celui de l'IP (donc résultat de soustraction négatif) annulent le signal des régions positives, le signal des régions négatives est ramené à zéro. L'approche d'Orlando a été sélectionnée pour le cœur de SpkNorm étant donné que sa simplicité permet de normaliser de façon efficace sans trop altérer les ensembles de données (pas de plancher de valeur) et permet une plus grande flexibilité d'usage. En effet, il est parfois utile de voir l'*input* et l'IP séparément, ce qui est possible avec cette approche. De plus, l'accès aux données IP normalisées, avant la soustraction de l'*input,* permet de faire d'autres manipulations, telles que normaliser en utilisant le signal d'un autre IP (e.g. normaliser les signaux de phosphorylation du CTD par le signal de l'ARNpol). Pour faciliter la comparaison de tous les ensembles de données sur la même échelle, un facteur de mise à l'échelle est ensuite calculé.

Bien qu'un facteur de mise à l'échèle puisse être calculé avec le nombre total de lectures alignées, spkNorm utilise la somme du signal. Cette alternative a été mise en place pour éviter un biais potentiel par la longueur des lectures qui peut varier entre les expériences. Ainsi, pour normaliser à une équivalence de 1M de lectures de 100 pb, il faut que la somme du signal dans des fenêtres de 10 pb soit égale à 10 M. En effet, la somme du signal (S_s) est égale à :

$$S_s = D_g * N_f$$

où D_g est la densité génomique et N_f est le nombre de fenêtres. La densité et le nombre de fenêtres peuvent être développés comme suit :

$$D_g = \frac{N_l * L_l}{L_g}$$

$$N_f = {L_g}/{L_f}$$

où :

$$N_l = Nombre\ de\ lectures$$

$$L_l = Longueur\ des\ lecures$$

$$L_g = Longueur\ du\ génome$$

$$L_f = Longueur\ des\ fenêtres$$

Alors la somme peut être développée et résolue comme suit :

$$S_s = D_g * N_f = \frac{N_l * L_l}{L_g} * \frac{L_g}{L_f} = \frac{N_l * L_l}{L_f} = \frac{1M * 100pb}{10pb} = 10M$$

Alors, pour faciliter la comparaison de tous les ensembles de données sur la même échelle, nous avons arbitrairement choisi de ramener le signal des souches WT de *S. pombe* à une somme de signal équivalant à 1 M de lectures de 100 bp fragmenté dans des fenêtres de 10 pb (soit une somme de signal total de 10 M). Exactement le même facteur de normalisation (calculé sur WT) est ensuite appliqué sur le Mut afin de ne pas changer le ratio entre le Mut et le WT. Ainsi, le WT sera toujours identique à la normalisation standard (et donc comparable avec des ensembles de données sans SI) puisque dans les deux cas, le signal total de *S. pombe* est ramené à 10 M.

Pour comparer la justesse entre l'application du SI et la normalisation standard fournis par l'outil SkpNorm, le profil agrégé d'un IP sur les gènes du SI a été effectué (Figure

8A). Il est attendu qu'une bonne normalisation doive faire en sorte que les profils de *S. cerevisiae* entre le WT et le Mut soient identiques puisque la chromatine ajoutée dans les deux échantillons est la même. Il est clair que le WT et le Mut normalisés avec le SI suivent une courbe similaire. Par contre, le Mut normalisé de façon standard est complètement décalé du WT. Ce résultat montre que la normalisation avec SI a bien été effectuée et que les résultats de ces expériences peuvent être interprétés avec confiance.

La Figure 8B compare les différentes conclusions biologiques interprétées à la suite de l'application du SI et la normalisation standard fournie par l'outil SkpNorm sur des données d'IP de l'ARNpol générées dans la souche WT et dans une souche Mut pour Ysh1, ayant un effet global de diminution de l'ARNpol II. Bien que la normalisation standard permette de mettre en évidence le défaut de terminaison (déplacement du sommet en 3'), l'option SI de SpkNorm met aussi en évidence la diminution généralisée l'ARNpol II dans le corps du gène.

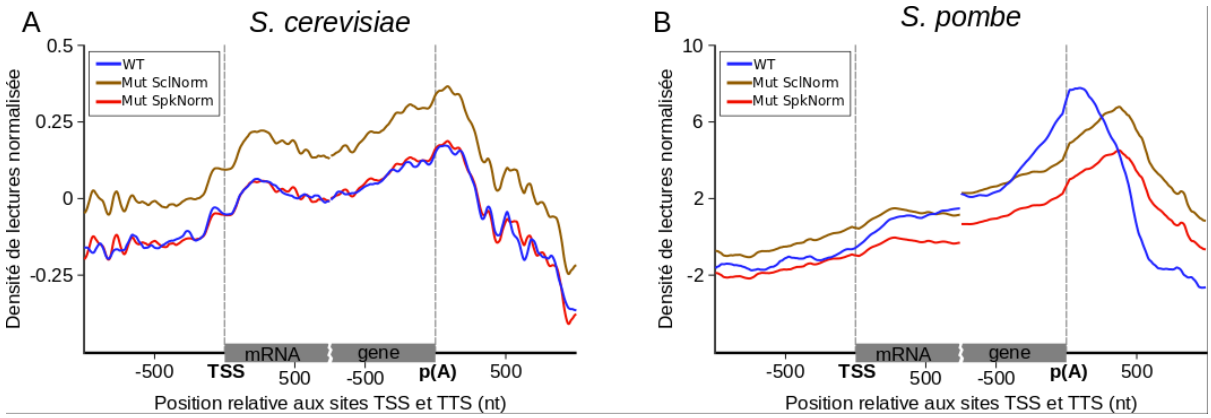


Figure 8. Comparaison du signal normalisé avec la normalisation standard et SpkNorm.

Profils agrégés de l'ARNpol II sur les gènes codants de *S. cerevisiae* (A) et *S. pombe* (B) pour un WT (bleu) et Mut pour Ysh1 normalisée avec la normalisation standard (SclNorm, brun) et la normalisation SpkNorm (rouge).

2.1.5. Étapes de SpkNorm

Voici les étapes de SpkNorm, après avoir aligné toutes les lectures d'une expérience avec SI sur un assemblage chimérique (contenant les deux génomes d'intérêt) et après avoir calculé la densité de signal dans des fenêtres de 10 pb en utilisant la moyenne pondérée :

a) Calculer le facteur de normalisation α de chaque ensemble de données en utilisant la totalité du signal aligné S sur le génome exogène (*S. cerevisiae*) rapportée sur 10 millions.

$$\alpha = \frac{1 * 10M}{S_{cerevisiae}}$$

b) Appliquer ce facteur sur le signal chaque fenêtre (S_f) des deux génomes pour obtenir le signal normalisé par le SI (S_{normSI_f}).

$$S_{normSI_f} = S_f * \alpha$$

c) Pour que tous les ensembles de données soient sur la même échelle ($sclWT$), calculer le facteur de mise à l'échelle des données WT (β) en utilisant seulement le signal provenant de la portion de *S. pombe* ($SWT_{normSI_{pombe}}$) pour chaque anticorps. Ce nouveau facteur est ensuite appliqué à chaque fenêtre des données WT (SWT) et Mut ($SMut$). IMPORTANT de ne pas recalculer le

deuxième facteur de mise à l'échelle avec les données Mut pour ne pas causer la perte de la première normalisation.

$$\beta = \frac{1 * 10M}{SWT_{normSi_{pombe}}}$$

$$SWT_{sclWT_f} = SWT_{normSi_f} * \beta$$

$$SMut_{sclWT_f} = SMut_{normSi_f} * \beta$$

d) La dernière étape consiste à soustraire dans chaque fenêtre le signal de l'*input* ($Sinput$) de l'immunoprécipitation (SIP). À noter que les données d'*input* auront aussi subi le traitement ci-haut.

$$SIP_{SpkNorm_f} = SIP_{sclWT_f} - Sinput_{sclWT_f}$$

2.2. Article décrivant le mécanisme commun de la terminaison de la transcription des gènes codants et non-codants chez la levure à fission

2.2.1. Introduction de l'article et contribution des auteurs

La terminaison de la transcription de l'ARNpol II est une étape complexe et fondamentale de l'expression des gènes qui est essentielle pour déterminer les frontières des gènes, mais qui peut aussi influencer la régulation de ceux-ci. Chez *S. cerevisiae*, la terminaison des gènes codants pour une protéine est initialisée par la machinerie de coupure et polyadénylation, tandis que la terminaison de la majorité des

gènes non-codants survient via une voie qui requière le complexe NNS. Des évidences supportent l'existence d'un complexe NNS non fonctionnel chez la levure à fission *S. pombe*. Bien que la terminaison soit plus étudiée chez *S. cerevisiae*, l'étude et la compréhension de ce mécanisme complexe chez *S. pombe* y sont complémentaires et tout aussi pertinentes pour permettre l'extrapolation vers l'humain. En effet, la distance évolutive entre les deux levures est presque aussi grande que celle entre ces levures et l'humain, formant ainsi un triangle évolutif.

Pour examiner la terminaison de la transcription de l'ARNpol II aux gènes non-codants chez *S. pombe*, nous avons analysé à l'échelle du génome la distribution des facteurs de terminaisons Rna14, Ysh1, Dhp1, Seb1, et Pcf11, puis nous avons comparé leur patron de recrutement aux marques de phosphorylation du CTD de l'ARNpol II. Nous démontrons que ces facteurs de terminaisons sont fortement recrutés en 3' des gènes non-codants, incluant les ARNsno et les ARNsn, et que leur recrutement coïncide avec de hauts niveaux de phosphorylation de la Ser2 et Tyr1 du CTD. De façon consistante avec le rôle des facteurs de terminaisons des ARNm, la déplétion de facteurs essentiels responsables de la coupure et polyadénylation a résulté en un défaut de terminaison de la production des ARNsno et ARNsn. En effet, il est montré que la terminaison des gènes non-codants est sensible à la déficience en Ysh1 et Dhp1, ce qui supporte le mécanisme coupure-dépendant de la terminaison de transcription. Ces données suggèrent un modèle pour lequel la terminaison de la transcription de l'ARNpol II chez *S. pombe* repose sur un mécanisme coupure-dépendant universel.

Marc Larochelle (M.L.) et François Bachand (F.B.) ont conçu l'étude et effectué les expériences de laboratoire, tandis que Marc-Antoine Robert (M-A.R.) et Pierre-Étienne Jacques (P-E.J.) ont planifié le traitement et les analyses des données de séquençages. M.L. a préparé les extraits de chromatine et a effectué les ChIP-Seq et autres expériences, incluant le contrôle qualité et la préparation de librairies. M.L. a

aussi fait la majorité des analyses ARN avec l'aide de Jean-Nicolas Hébert (J-N.H.). J-N.H. a fait les constructions Ysh1 et a performé la caractérisation phénotypique des Mut *anchor-away ysh1* et *rna14.* Xiaochuan Liu (X.L.) a préparé les librairies pour les données de *3' region extraction and deep sequencing* (3'READS) avec l'aide de Bin Tian (B.T). Dominick Matteau (D.M.) et Sébastien Rodrigue (S.R.) ont aidé avec la préparation des librairies des ChIP-Seq. M.L., M-A.R., P-E.J. et F.B. ont préparé et finalisé les figures. F.B. a écrit le manuscrit, avec l'aide de M.L., M-A.R. et P-E.J.

Référence bibliographique : Larochelle, M., Robert, M-A., Hébert, J-N., Lui, X., Matteau, D., Robrigue, S., Tian, B., Jacques, P-É. and Bachand, F. (2018). Common mechanism of transcription termination at coding and noncoding RNA genes in fission yeast. En révision à Nature Communications.

Common mechanism of transcription termination at coding and noncoding RNA genes in fission yeast

Marc Larochelle[1]†, Marc-Antoine Robert[2]†, Jean-Nicolas Hébert[1], Xiaochuan Liu[3], Dominick Matteau[2], Sébastien Rodrigue[2], Bin Tian[3], Pierre-Étienne Jacques[2,4]* & François Bachand[1,4]*

1. RNA Group, Département de Biochimie, Université de Sherbrooke, Sherbrooke, Quebec, Canada
2. Départment de Biologie, Université de Sherbrooke, Sherbrooke, Quebec, Canada
3. Department of Microbiology, Biochemistry and Molecular Genetics, Rutgers New Jersey Medical School and Rutgers Cancer Institute of New Jersey, Newark, New Jersey, USA
4. Centre de Recherche du CHUS, Université de Sherbrooke, Sherbrooke, Québec, Canada

†These authors contributed equally to this work.

*Correspondence should be addressed to P-E. J. (Pierre-Etienne.Jacques@USherbrooke.ca) and F.B. (f.bachand@usherbrooke.ca).

Running Title: *A universal mode of transcription termination in fission yeast*

2.2.2. Abstract

Termination of RNA polymerase II (RNAPII) transcription is a fundamental step of gene expression that is critical for determining the borders between genes. In budding yeast, termination at protein-coding genes is initiated by the cleavage/polyadenylation machinery, whereas termination of most noncoding RNA (ncRNA) genes occurs via the Nrd1-Nab3-Sen1 (NNS) pathway. Unexpectedly, we show here that NNS-like transcription termination is not conserved in fission yeast. Instead, genome-wide location analyses show global recruitment of mRNA 3' end processing factors at the end of ncRNA genes, including snoRNAs and snRNAs, and that this recruitment coincides with high levels of Ser2 and Tyr1 phosphorylation on the RNAPII C-terminal domain. We further show that termination of mRNA and ncRNA transcription requires the conserved Ysh1/CPSF-73 and Dhp1/XRN2 nucleases, supporting widespread cleavage-dependent transcription termination in fission yeast. Our findings thus reveal that a common mode of transcription termination can produce functionally and structurally distinct types of polyadenylated and non-polyadenylated RNAs.

2.2.3. Introduction

In eukaryotes, RNA polymerase II (RNAPII) is responsible for the synthesis of a broad range of coding and noncoding transcripts, and termination pathways have a decisive influence on the fate of transcribed RNAs[1]. Although transcription termination has been investigated in a wide range of organisms, it has been most extensively studied in the model organism *Saccharomyces cerevisiae* where two major termination pathways exist depending on the class of genes transcribed[2]. For protein-coding genes, existing data support a mechanism triggered by the cleavage activity of the Ysh1 endonuclease (CPSF-73 in humans), which is part of a nuclease module within the larger mRNA 3'

end processing complex[3] Specifically, the co-transcriptional recruitment of conserved cleavage and polyadenylation factors at poly(A) signal (PAS) causes cleavage of the nascent pre-mRNA, followed by 3' end polyadenylation of the released transcript by the poly(A) polymerase[4]. The endonucleolytic cleavage also provides a free and uncapped 5' entry point for a protein complex that includes an evolutionarily conserved 5'-3' exonuclease (XRN2 in humans, Rat1 in *S. cerevisiae;* Dhp1 in *S. pombe*). The exonuclease is thought to chase RNAPII and promote its dissociation from the DNA template[5-8], a mechanism referred as 'torpedo'-mediated transcription termination.

In addition to mRNAs, RNAPII also synthesizes an extensive set of noncoding RNAs (ncRNAs) that include small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), long noncoding RNAs (lncRNAs), and cryptic unstable transcripts (CUTs). In *S. cerevisiae*, transcription of snoRNAs, snRNAs, and CUTs does not rely on the cleavage/polyadenylation machinery for termination, but instead uses a cleavage-independent termination pathway that requires a complex (referred to as NNS complex) consisting of the RNA-binding proteins Nab3 and Nrd1 as well as the DNA/RNA helicase Sen1[2]. In this mode of termination, the NNS complex interacts with both the transcription machinery[9,10] and specific RNA motifs enriched downstream of ncRNA genes[11,12] to engage the transcription elongation complex via the Sen1 helicase, which translocate onto the nascent RNA and catches up with the transcribing polymerase to elicit termination[13]. NNS-dependent termination is coupled to the addition of short poly(A) tails by the TRAMP polyadenylation complex, which targets ncRNAs for 3' end maturation by exonucleolytic trimming (e.g. snoRNAs) or complete degradation (e.g. CUTs) by the exosome complex of 3'-5' exonucleases[14,15].

One transcription-associated feature that appears to influence the choice between torpedo- and NNS-mediated termination is the phosphorylation status of the carboxy-terminal domain (CTD) of the RNAPII catalytic subunit, Rpb1. The CTD consists of a

succession of conserved heptad repeats, with the consensus amino acid sequence Y-S-P-T-S-P-S[16]. The RNAPII CTD is subjected to a plethora of post-translational modifications throughout the transcription cycle that are key to coordinate the sequential recruitment of RNA processing factors. Among CTD modifications, phosphorylation of Ser2 and Ser5 have been most extensively studied and appear to be the most abundant in *S. cerevisiae*[17]. In the cases of small ncRNA genes, NNS recruitment is influenced by Ser5 phosphorylation (Ser5-P) via the CTD-interaction domain (CID) of Nrd1[9,10]. As transcription elongation progresses into protein-coding genes, the ratio between the levels of phosphorylated Ser2 (Ser2-P) and Ser5-P gradually increases in the RNAPII CTD, peaking near the mRNA cleavage/poly(A) site[18,19]. Pcf11, a component of the mRNA 3' end processing machinery, preferentially recognizes Ser2-P CTD repeats via its CID domain[20,21], which may explain the robust levels of Pcf11 at the 3' end of protein-coding genes[22]. Yet, Pcf11 recruitment at the 3' end of mRNA genes may also be influenced by additional CTD modifications. For instance, Tyr1 phosphorylation (Tyr1-P) can impair binding of Pcf11 to Ser2-P CTD peptides *in vitro*[22]. Accordingly, it was proposed that Tyr1-P prevents recruitment of termination factors in the coding region of genes in *S. cerevisiae*, and that timely dephosphorylation of Tyr1-P upstream of the polyadenylation site by the Glc7 phosphatase would allow for recruitment of the 3' end processing machinery via recognition of Ser2-P CTD repeats by the CID domain of Pcf11[23].

High-throughput sequencing analyses indicate that the majority of genomes are transcriptionally active, thus yielding a large amount of ncRNAs[24,25]. Yet, how RNAPII transcription is terminated at ncRNA genes remains poorly understood for many eukaryotic species. Also unclear is whether the use of distinct pathways to terminate coding and noncoding transcription, such as described in *S. cerevisiae*, is evolutionarily conserved. Recently, we have shown that the *S. pombe* homolog of *S. cerevisiae* Nrd1, Seb1, does not physically associate with Nab3 and Sen1 homologs in fission yeast[26]. Here we show that the absence of Nab3 and Sen1 homologs in *S. pombe* does not

affect transcription termination of coding and ncRNA genes. Unexpectedly, we found that mRNA 3' end processing factors are enriched at the 3' end of ncRNA genes and that this recruitment coincides with high levels of Ser2 and Tyr1 CTD phosphorylation. Consistently, we find that most independently-transcribed fission yeast snoRNA genes are cleaved and polyadenylated in a manner dependent on conserved 3' end processing factors, and that termination at ncRNA genes is sensitive to deficiencies in Ysh1 and Dhp1. Our findings indicate that torpedo-mediated transcription termination is widespread in fission yeast, thereby revealing that a universal mode of transcription termination can promote the synthesis and accumulation of both polyadenylated mRNAs and non-polyadenylated ncRNAs.

2.2.4. Results

2.2.4.1. Absence of RNAPII termination defects in *S. pombe nab3Δ* and *sen1Δ* mutants

We have recently shown that Seb1, the sequence homolog of *S. cerevisiae* Nrd1, does not assemble into a stable complex with Nab3 and Sen1 proteins in *S. pombe* [26]. Furthermore, analysis of RNAPII transcription at a model snoRNA gene in fission yeast (*snR3*) did not reveal transcription termination defects in *nab3Δ* and *sen1Δ* mutants [26]. To measure the global impact of mutations in NNS components on RNAPII transcription in *S. pombe*, we analyzed genome-wide RNAPII (Rpb1) occupancy by ChIP-seq using *nab3Δ*, *sen1Δ*, and *dbl8Δ* (Sen1 paralog) single mutants as well as a *sen1Δ dbl8Δ* double mutant. As shown in Fig. 9A- 9B, the absence of Nab3, Sen1, and Dbl8 did not result in noticeable read-through transcription at mRNA and snoRNA genes compared to the wild-type (WT) control. In striking contrast, a deficiency in the essential protein Seb1 results in defective termination at most RNAPII-transcribed genes (see P_{nmt1}-

seb1 mutant; Fig. 9A- 9B) [26]. Given that Nab3, Sen1, and Dbl8 do not copurify with Seb1 [26,27], and considering the absence of termination defects in *nab3Δ, dbl8Δ*, *sen1Δ*, and *sen1Δ dbl8Δ* mutants (Fig. 9A-9B), we conclude that NNS-dependent transcription termination is not conserved in fission yeast.

2.2.4.2. mRNA 3' end processing factors are recruited at the 3' end of coding and noncoding RNAPII-transcribed genes

The surprising lack of conservation of NNS-dependent transcription termination raised the question as to the mechanism of 3' end processing and transcription termination at ncRNA genes in fission yeast. One clue about how ncRNAs could be processed in *S. pombe* derives from studies on the nuclear poly(A)-binding protein Pab2, the homolog of human PABPN1 [28]. Our work previously uncovered a polyadenylation-dependent pathway required for the maturation of independently-transcribed snoRNAs that relies on Pab2 [29,30]. However, the mechanism responsible for 3' end processing and transcription termination had remained elusive. To test for the possibility that mRNA cleavage and polyadenylation factors function in 3' end processing of ncRNA genes in *S. pombe*, we used ChIP-seq to examine the genome-wide binding profile of proteins with conserved roles in mRNA 3' end processing and transcription termination. This analysis included two independent components of the cleavage and polyadenylation factor I (CFI) complex, namely Rna14 and Pcf11, the endonuclease Ysh1 that is responsible for pre-mRNA cleavage before 3' end polyadenylation, and Dhp1, the homolog of *S. cerevisiae* Rat1 and human XRN2 that promote transcription termination of RNAPII via 5'-3' exonucleolytic activity. The ChIP-seq profile of Seb1 [26] was also included, as this protein was shown to be important for 3' end processing of both mRNAs and ncRNAs. Consistent with roles in 3' end processing at mRNA genes, Rna14, Pcf11, Ysh1, Seb1, and Dhp1 all showed strong binding at the 3' end of protein-coding genes (Fig. 9C and 9F).

Notably, they also showed robust enrichment at the 3' end of independently-transcribed snoRNAs (Fig. 9D and 9G) and snRNA (Fig. 9E) genes. In contrast, these factors were not enriched at the 3' end of intron-encoded snoRNAs (Fig. S1), consistent with a maturation pathway independent of cleavage and polyadenylation. Interestingly, analysis of peak distribution based on average binding profiles indicated that the recruitment of Dhp1 generally occurred downstream of the endonuclease Ysh1 (Fig. 9F-9G; note that the peak of the green curves is shifted downstream relative to the pink curves). This is consistent with transcription termination of coding and ncRNA genes following the "torpedo" model, in which Dhp1 requires Ysh1-dependent endonucleolytic cleavage to promote disengagement of RNAPII from the DNA template [5-7]. On the basis of these results, we conclude that mRNA 3' end processing factors are recruited at the 3' end of both coding and ncRNA genes in fission yeast.

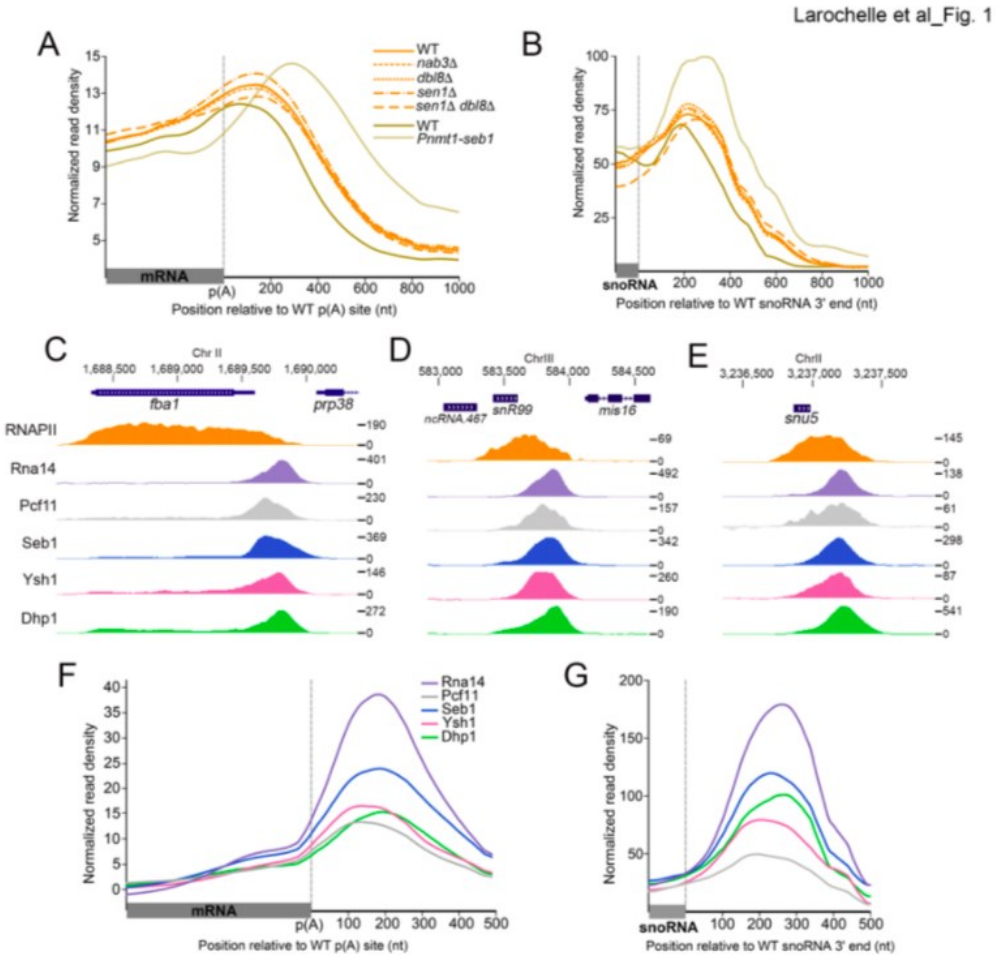


Figure 9. *S. pombe* mRNA 3' end processing and transcription termination factors are recruited at the 3' end of independently-transcribed snoRNA and snRNA genes.

(A-B) Average ChIP-seq profiles of total RNAPII (Rpb1) relative to mRNA poly(A) site (A, n=4,755) and to annotated 3' end of independently-transcribed monocistronic snoRNAs (B, n=31) in WT (solid line, orange) and NNS mutant strains (dotted lines, orange) grown in rich medium. Average ChIP-seq profiles of total RNAPII (Rpb1) in WT and Seb1-deficient cells (*Pnmt1-seb1*) grown in thiamine-supplemented minimal medium are also shown (gold). Gene coordinates are available in Table S2-S3. (C-E) Normalized ChIP-seq signal of RNAPII (Rpb1) and the indicated mRNA 3' end processing factors across the *fba1* mRNA (C), the *snR99* snoRNA (D), and the *snu5* snRNA (E) genes. (F-G) Average ChIP-seq profile of the indicated mRNA 3' end processing factors over the same groups of mRNA (F) and snoRNA (G) genes as panels A-B.

2.2.4.3. mRNA 3' end processing factors are required for the synthesis of ncRNAs

The presence of cleavage and polyadenylation factors at the 3' end of ncRNA genes suggested a role for the mRNA 3' end processing machinery in snoRNA synthesis. To examine the requirement of Rna14, Pcf11, Ysh1, and Dhp1 in the synthesis of ncRNAs, conditional strains were generated since these proteins are all encoded by essential genes. As used previously to repress the expression of essential core exosome subunits [31] and Seb1 [26], we constructed strains with endogenous *pcf11* and *dhp1* genes under the control of the thiamine-repressible *nmt1* promoter (P_{nmt1}). Consistent with Pcf11 and Dhp1 being essential for viability [32], P_{nmt1}-*pcf11* and P_{nmt1}-*dhp1* cells cultured in thiamine-supplemented medium showed growth arrest (Fig. S2A). Depletion of Pcf11 in *S. pombe* also affects mRNA production [33], as expected of a protein essential for mRNA synthesis. To generate conditional strains for *rna14* and *ysh1*, we used the rapamycin-dependent anchor-away system (Fig. S2B) [34], as the *nmt1* promoter did not provide sufficient Rna14 and Ysh1 depletion to induce growth arrest (data not shown). Importantly, inactivation of Ysh1 and Rna14 by nuclear depletion (Fig. S2C-S2D) impaired mRNA synthesis in a rapamycin-dependent manner (Fig. S2E-S2F) and resulted in the production of read-through transcripts (Fig. S2G), consistent with 3' end processing defects at protein-coding genes. To assess the contribution of Pcf11, Dhp1, Ysh1, and Rna14 to snoRNA 3' end processing, we used an RNase H cleavage assay that can simultaneously detect both mature snoRNAs and 3'-extended polyadenylated snoRNA precursors (Fig. 2A). As a control, we used the *pab2Δ* mutant that accumulates 3'-extended polyadenylated precursors, resulting in reduced levels of mature snoRNAs [30] (Fig. 10B-10C, compare lanes 3-4 to 1-2). Deficiencies in Pcf11, Ysh1, and Rna14 resulted in reduced levels of mature snoRNAs (Pcf11: see Fig. 10B-10C, lanes 5-6; Ysh1 and Rna14: see Fig. 10D-10E, lanes 3-6). In addition to polyadenylated pre-snoRNA detected by RNase H assays (Fig. 10B-10E), long *snR3* read-through products were also detected by Northern blot analysis in cells deficient for Pcf11, Ysh1, and Rna14 (Fig. 10F, lanes 2-3 and 5). In contrast, snoRNA 3'-

extensions did not accumulate in cells deficient for Dhp1 (Fig. 10B-10C, lanes 9-10 and Fig. 10F, lane 7), consistent with a role in transcription termination that occurs beyond 3' end processing. As shown previously [26], a deficiency in Seb1 did not negatively impair RNA production as demonstrated by the similar levels of mature snoRNA between wild-type and Seb1-deficient cells (Fig. 10B-10C, compare lanes 1-2 and 7-8), but changed cleavage site selection as observed by the lengthening of the 3'-extended snoRNA precursors on RNase H assays. These results indicate that mRNA 3' end processing factors are required for the synthesis of independently-transcribed snoRNAs. Together with data showing recruitment of 3' end processing factors downstream of ncRNA genes (Fig. 9), our results suggest that ncRNA 3' end maturation involves cleavage and polyadenylation.

2.2.4.4. Widespread polyadenylation of independently-transcribed snoRNAs

Our current and previous work on polyadenylation of pre-snoRNAs has focused mainly on a handful of snoRNA genes [29,30]. To obtain a more comprehensive view of pre-snoRNA polyadenylation and to address whether differences exist between 3' end processing of H/ACA and C/D box snoRNAs, we used a sensitive sequencing approach since polyadenylated pre-snoRNAs are rapidly processed; thus being present at low levels in wild-type cells as compared to the stable, non-polyadenylated mature snoRNAs (Fig. 10B-10E, lanes 1-2). We therefore used 3' Region Extraction And Deep Sequencing (3'READS), an approach developed to map mRNA polyadenylation sites at the genome-wide level [35].

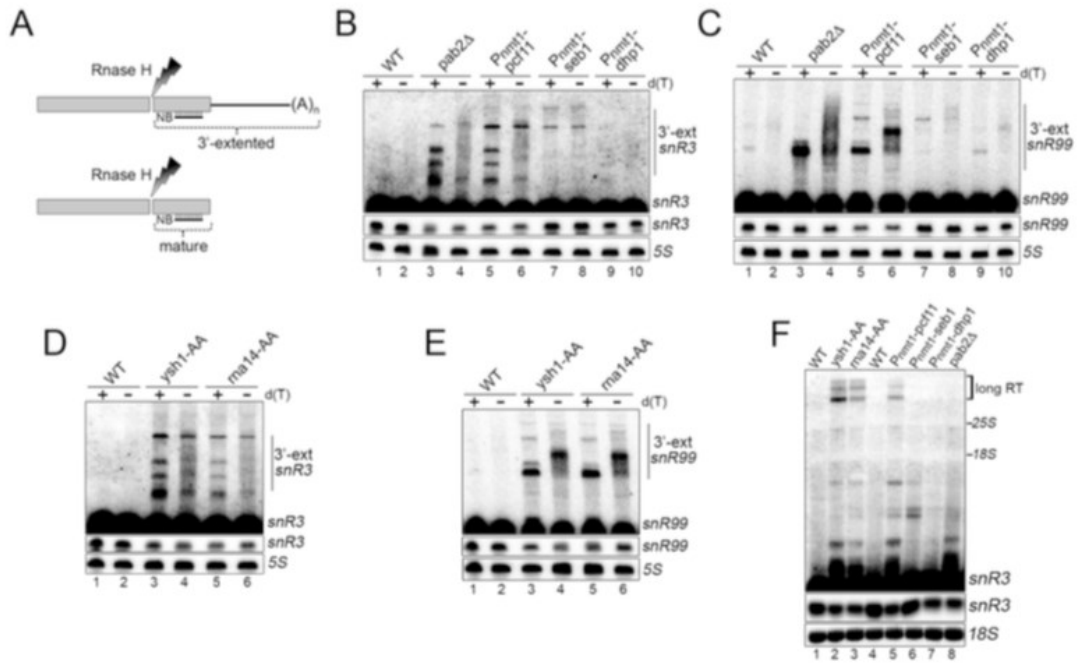


Figure 10. The mRNA cleavage and polyadenylation complex is required for snoRNA synthesis.

(A) Schematic of the RNase H cleavage assay used in panels B-E. After RNase H cleavage of the snoRNA at sites of RNA:DNA hybrids in the presence of a sequence-specific DNA oligonucleotide, the 3' fragment (mature or 3'-extended) is detected by Northern blotting (NB). Addition of oligo d(T) to the RNase H reaction removes heterogenous poly(A) tails, generating discrete products. (B-C) Total RNA prepared from the indicated strains that were previously grown in thiamine-supplemented medium to deplete Pcf11, Seb1, and Dhp1 was treated with RNase H in the presence of DNA oligonucleotides complementary to *snR3* (B) and *snR99* (C). RNase H reactions were performed in the presence (+) or absence (-) of oligo d(T). The top panel represents a longer exposure of the middle panel to see 3'-extended (3'-ext) cleavage products. The 5S rRNA was used as a loading control. (D-E) As described in B-C, but using cells that were previously treated with rapamycin to deplete Ysh1 and Rna14 from the nucleus. (F) Northern blot analysis using total RNA prepared from the indicated strains that were treated with either rapamycin (lanes 1-3) or thiamine (lanes 4-8). The blot was hybridized using DNA probes specific to *snR3* and the 18S rRNA. The position of mature *snR3*, 18S and 25S rRNAs, as well as *snR3* read-through (RT) products is indicated on the right.

Over 6 millions reads mapping to polyadenylation sites (PAS reads) were obtained from wild-type *S. pombe* cells grown in rich and minimal media by 3'READS [36]. Although the majority (~60%) of unique PAS reads mapped to 3' UTR of protein-coding genes, roughly 18% of unique PAS locations mapped to intergenic regions. Notably, out of 31 independently-transcribed monocistronic snoRNAs, 24 (77%) and 27 (87%) had a mappable poly(A) site located downstream of a snoRNA annotation in minimal and rich media, respectively (Table S3). It should be noted that snoRNA genes for which PAS reads could not be identified showed barely detectable levels of RNAPII as determined by ChIP-seq assays (data not shown). Comparison between C/D and H/ACA box snoRNAs indicated similar proportions of polyadenylated species (Fig. 11A). Figure 11B shows 3'READS results for the H/ACA box *snR99* snoRNA. As can be seen, 3'READS identified a major PAS located 78-nt downstream of the annotated *snR99* mature 3' end in wild-type cells. In the *pab2Δ* mutant, 3'READS detected a >50-fold increase in the abundance of *snR99* precursors that used this major PAS (Fig. 11B). Notably, whereas depletion of Seb1 resulted in the use of a single distal PAS located 118-nt downstream from the major *snR99* PAS (196-nt from the *snR99* mature 3' end), the two different PASs identified in the WT, *pab2Δ* and *Pnmt1-seb1* strains were used in cells deficient for Pcf11 (Fig. 11B). Collectively, the 3'READS data are entirely consistent with results obtained using RNase H cleavage assays (see Fig. 10C).

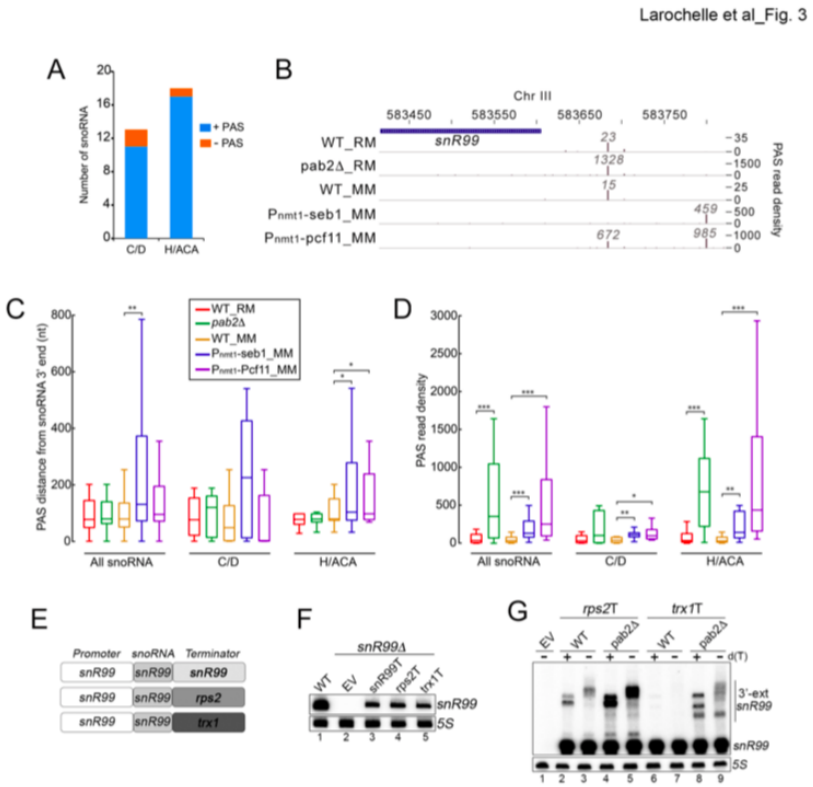


Figure 11. Independently-transcribed snoRNA genes are cleaved and polyadenylated.

(A) Proportion of 13 C/D box and 18 H/ACA box monocistronic snoRNAs with at least one poly(A) site mapped by 3'READS in the WT strain grown in minimal or rich media. (B) Poly(A) site (PAS) read density profile downstream of the *snR99* snoRNA as determined by 3'READS in the indicated strains grown in either rich (RM) or minimal (MM) media supplemented with thiamine to deplete Seb1 and Pcf11. (C-D) Distribution of the distance calculated between the strongest poly(A) site (PAS) and the annotated snoRNA 3' end (C), as well as the sum of the read density for all of the poly(A) sites associated to a snoRNA in each condition (D) (* pval<0.05; ** pval<0.01; *** pval<0.005, Wilcoxon signed-rank test). (E) Schematic of *snR99* constructs with different 3' flanking sequences used for experiments presented in panels F-G. (F) Northern blot analysis of total RNA prepared from WT (lane 1) and *snR99*-null (lanes 2-5) cells that were previously transformed with the indicated constructs (EV, empty vector). (G) Total RNA prepared from *snR99*-null cells that were previously transformed with *snR99* constructs that comprised *rps2* (rps2T, lanes 2-5) or *trx1* (trx1T, lanes 6-9) terminator sequences was treated with RNase H in the presence of DNA oligonucleotides complementary to *snR99*.

Global analysis of 3'READS data revealed that the median distance between the annotated (mature) snoRNA 3' end and the pre-snoRNA cleavage site is 77-nt in wild-type cells grown in rich medium (Fig. 11C, red boxes). Consistent with previous results [30] (Fig. 10), the absence of Pab2 did not generally affect cleavage site selection at snoRNA genes (Fig. 11C, green boxes), but resulted in a significant accumulation of polyadenylated snoRNA precursors (Fig. 11D, compare green and red boxes); interestingly, H/ACA box pre-snoRNAs appeared to accumulate to greater levels as compared to C/D box precursors in the absence of Pab2 (Fig. 11D). In contrast, the absence of Pab2 did not generally impact the levels of polyadenylated mRNAs (Fig. S3B). In the case of Seb1, as seen for *snR99* (Fig. 11B), we noted a clear shift of the general distribution of polyadenylated pre-snoRNAs towards longer 3'-extensions (Fig. 11C, blue boxes): the median distance between the snoRNA mature 3' end and its corresponding poly(A) site increased from 79-nt in wild-type cells grown in minimal medium to 131-nt in Seb1-deficient cells (compare blue and yellow boxes). This result is consistent with the preferential use of distal cleavage sites in the *seb1* mutant (Fig. 10), an outcome also observed for mRNAs [26] (Fig. S3A). The overall distribution of pre-snoRNA polyadenylation sites was also extended in Pcf11-deficient cells, a result that was greater for H/ACA box snoRNAs (Fig. 11C, compare purple and yellow boxes), as well as for mRNAs (Fig. S3A) [36]. In addition, a general accumulation of polyadenylated pre-snoRNAs was observed in *pcf11* and *seb1* mutants (Fig. 11D, compare purple and blue boxes to yellow boxes). In summary, results presented in Figure 11 indicate that most independently-transcribed snoRNA genes produce 3'-extended polyadenylated precursors in *S. pombe*.

A mechanism of snoRNA 3' end processing that involves cleavage and polyadenylation by mRNA maturation factors posits that *cis*-acting elements present in the 3' flanking region of coding and ncRNA genes should be similar. We and others have shown that poly(A) signals around *S. pombe* mRNAs are associated with upstream AAU[A/G]AA hexamers and downstream GUA motifs [36-38]. However, the relatively small number of

independently-transcribed snoRNAs (n=31), as compared to mRNAs (n=4,755), limited the identification of significantly enriched sequence motifs around snoRNA poly(A) sites. We therefore tested whether *cis* elements promoting mRNA 3' end processing could support 3' end maturation of a snoRNA. For this, we generated constructs in which 1-kb of *snR99* downstream sequence was exchanged with 1-kb of downstream sequence from *rps2* and *trx1* protein-coding genes (Fig. 11E). These three constructs as well as a vector control were chromosomally integrated as a single copy into a *snR99*-null strain. As shown in Fig. 11F, 3' flanking sequences from both *rps2* and *trx1* supported proper 3' end formation and normal *snR99* accumulation, as demonstrated by the length and expression level of the snoRNA, which were similar to *snR99* expressed from its authentic downstream sequence (compare lanes 3-5). Furthermore, RNase H cleavage assays and 3' RACE experiments revealed Pab2-dependent polyadenylated snoRNA precursors that used the *rps2* and *trx1* cleavage sites (Fig. 11G and data not shown). Based on these findings, we conclude that snoRNA 3' end processing in *S. pombe* occurs via a pathway analogous to mRNA biogenesis.

2.2.4.5. Tyrosine 1- and Serine 2-phosphorylated forms of the RNAPII CTD colocalize with 3' end processing factors at coding and ncRNA genes

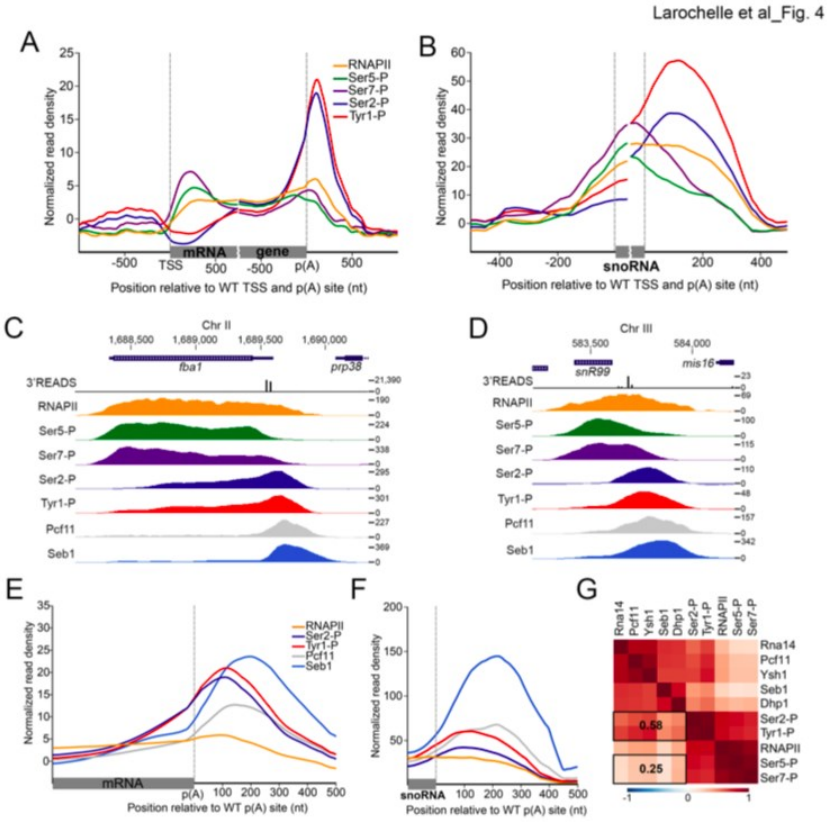
A number of studies have examined the association between the status of CTD phosphorylation and 3' end processing/transcription termination. For instance, in *S. cerevisiae* and humans, Ser2 phosphorylation occurs during transcription elongation and promotes the recruitment of factors that function in 3' end processing and termination [16,39]. Since little is known about the association between CTD modifications and RNA processing in fission yeast, we mapped Tyr1, Ser2, Ser5, and Ser7 phosphorylation at the genome-wide level in wild-type *S. pombe*. ChIP-seq data from independent biological replicates showed high correlation throughout the *S. pombe* genome (Fig. S4). Averaging ChIP-seq signal across protein-coding genes revealed

CTD phosphorylation patterns similar to those characterized in *S. cerevisiae* (Fig. 12A): Ser5 and Ser7 phosphorylation (Ser5-P and Ser7-P) marks peaking at the 5' end of coding genes after the transcription start site (TSS), whereas Ser2 phosphorylation (Ser2-P) steadily increased along the gene body and sharply peaked downstream of the poly(A) site. The pattern of Tyr1 CTD phosphorylation (Tyr1-P) at protein-coding genes was similar to that of Ser2-P, peaking sharply downstream of the poly(A) site (Fig. 12A). It is interesting to note that the genome-wide pattern of Tyr1-P signal in *S. pombe* contrasts to that of *S. cerevisiae* in which Tyr1-P signal was shown to peak upstream of the poly(A) site [22]. Occupancy profiles of CTD phosphorylation marks at snoRNA genes showed a pattern somewhat similar to that of mRNA genes, with rapid bursts of Ser5-P and Ser7-P signals at the 5' end that ultimately started decreasing in the snoRNA gene body (Fig. 12B). As seen for mRNA genes, Ser2-P and Tyr1-P signal steadily increased along the body of snoRNA genes, with peaks that coincided with the decline in total RNAPII levels downstream of snoRNA poly(A) sites (Fig. 12B). Together, our findings indicate that termination of RNAPII transcription at mRNA and snoRNA genes in *S. pombe* occurs at regions where maximal signal of Ser2 and Tyr1 phosphorylation are detected.

Pcf11 and Seb1 have CTD-interacting domains (CID) that preferentially recognize Ser2-phosphorylated RNAPII *in vitro* [21,27]. Accordingly, we compared the ChIP-seq profiles of Pcf11 and Seb1 to genome-wide RNAPII Ser2-P signal. As shown for a protein-coding and a snoRNA gene (Fig. 12C-12D), the peak occupancy of Pcf11 and Seb1 binding generally coincided with regions of maximum Ser2-P signal that were found downstream of poly(A) sites. The average occupancy profiles of Pcf11 and Seb1 also colocalized with regions of maximal Tyr1-P signal (Fig. 12C-12F). Globally, ChIP-seq signals of 3' end processing factors were found to be more correlated genome-wide with Ser2 and Tyr1 phosphorylation than with Ser5 and Ser7 phosphorylation (Fig. 12G, 0.58 vs 0.25). These results support the view that Ser2 and Tyr1 CTD phosphorylation are functionally relevant to fission yeast 3' end processing and

transcription termination. Furthermore, our data suggest that Tyr1 phosphorylation contributes differently to termination of RNAPII transcription between budding and fission yeasts.

Figure 12. Tyr1-P and Ser2-P forms of the RNAPII CTD colocalize with 3' end processing factors at coding and ncRNA genes.



(A-B) Average ChIP-seq profile of total RNAPII (Rpb1) and the indicated CTD modifications in a WT strain across 4,755 mRNAs (A) and 24 monocistronic snoRNAs (B) with a mapped poly(A) site in minimal medium. (C-D) PAS read density (3'READS) and normalized ChIP-seq signal of total RNAPII (Rpb1), the indicated CTD modifications, as well as Pcf11 and Seb1 across the *fba1* mRNA (C) and the *snR99* snoRNA (D) genes. (E-F) Average profile of total RNAPII (Rpb1), Ser2-P, Tyr1-P, Pcf11, and Seb1 relative to the poly(A) site of 4,755 mRNA (E) and 24 monocistronic snoRNA (F) genes in minimal medium. (G) Genome-wide pairwise Pearson correlation coefficient matrix at a resolution of 10 bp followed by a hierarchical clustering. Rectangles include the average of a subset of correlation values between 3' processing factors and CTD modifications.

2.2.4.6. "Torpedo" exonuclease-dependent RNAPII release is the general mode of transcription termination in fission yeast

Our results suggest that a common pathway that involves cleavage and polyadenylation promotes transcription termination at both coding and ncRNA genes in fission yeast. Cleavage-dependent transcription termination depends on the 5'-3' exonucleolytic activity of the conserved XRN2 nuclease (Dhp1 in *S. pombe*). To address whether termination of snoRNA transcription in *S. pombe* was generally dependent on Dhp1, we analyzed RNAPII occupancy in Dhp1-depleted cells. Because deficiencies in Rat1 and XRN2 have been shown to affect RNAPII occupancy [6,40], we introduced a spike-in normalization step [41] using a constant amount of *S. cerevisiae* chromatin to directly compare ChIP-seq samples between wild-type and Dhp1-depleted cells (see supplementary methods for details). As shown in Fig. 13A-13B, the distribution of total RNAPII (Rpb1) displayed increased density at the 3' end of the *fba1* mRNA and the *snR99* snoRNA in Dhp1-deficient cells, consistent with read-through transcription. Evidence of delayed transcription termination in the *dhp1* mutant was also noted at snRNA genes, as indicated by a shift of Rpb1 density into the 3' flanking region of *U5* and *U1* snRNAs (Fig. 13C and Fig. S5A). Read-through transcription was also detected by a CTD-independent ChIP approach that used an antibody to the HA-tag of a core RNAPII component (Rpb3-HA; Fig. 13A-13C). Importantly, transcription termination defects are a general feature of Dhp1-deficient cells, as seen by a clear shift in the Rpb1 and Rpb3-HA average signals downstream of the noticeable decline observed in wild-type cells (Fig. 13D-13E; compare solid and dotted lines). These results show that Dhp1 is required for RNAPII termination at both coding and ncRNA genes.

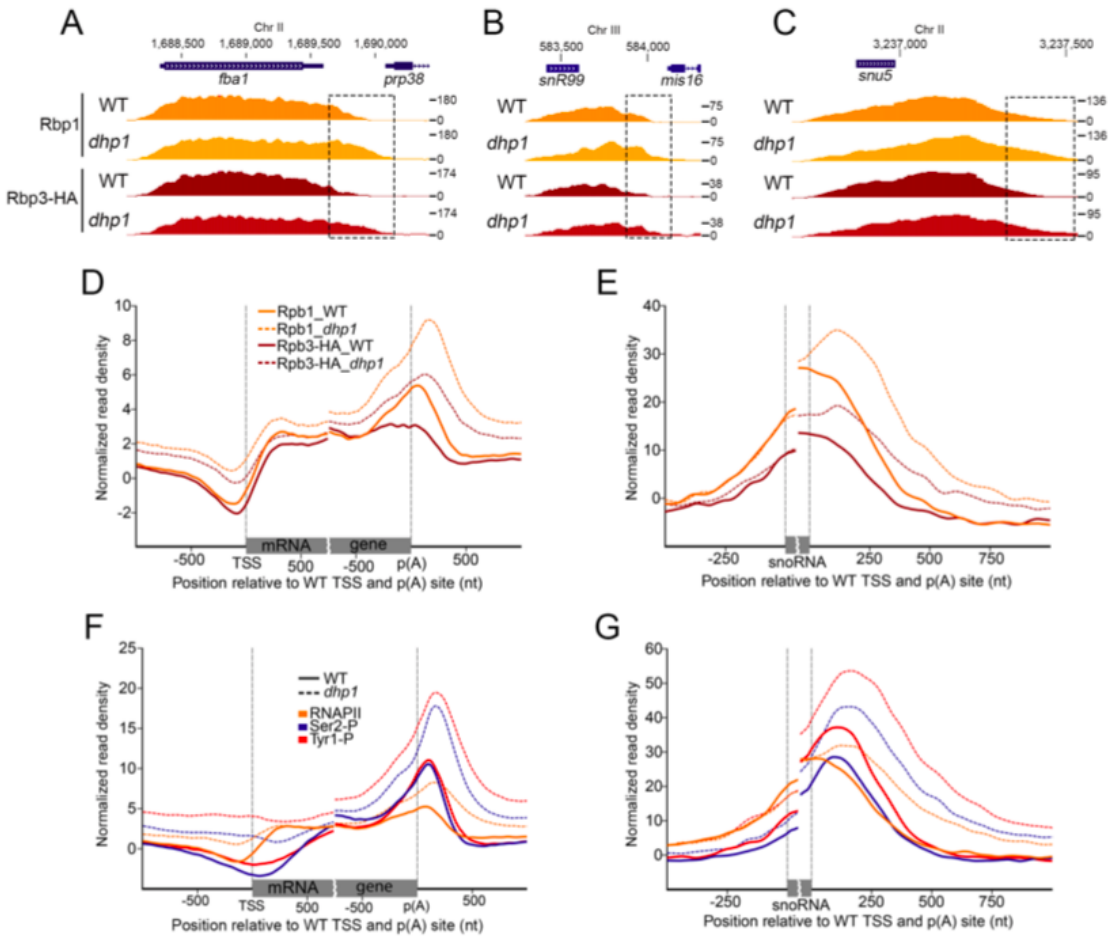


Figure 13. The torpedo nuclease Dhp1 is required for transcription termination of coding and ncRNA genes.

(A-C) Normalized ChIP-seq signal of RNAPII subunits Rbp1 and Rbp3 in WT and Dhp1-depleted strains across the *fba1* mRNA (A), the *snR99* snoRNA (B), and the *snu5* snRNA (C) genes in thiamine-treated minimal medium. The dashed-line rectangles highlight delayed transcription termination in Dhp1-deficient cells. (D-G) Average ChIP-seq profile of Rbp1 and Rbp3 (D-E) or total RNAPII (Rbp1), Ser2-P, and Tyr1-P (F-G) in WT (solid lines) or Dhp1-depleted (dotted lines) cells across 4,755 mRNA (D, F) and 24 monocistronic snoRNA (E, G) genes with a mapped p(A) site in thiamine-treated minimal medium.

Hyperphosphorylation of CTD repeats on Ser2 was previously observed in a temperature-sensitive mutant of *S. cerevisiae rat1* [40]. To address how Dhp1 inactivation globally affects CTD phosphorylation at transcribed *S. pombe* genes, we compared Tyr1, Ser2, Ser5, and Ser7 phosphorylation levels between Dhp1-deficient and control cells using spike-in-normalized ChIP-seq. As shown in Fig. 13F-13G, inactivation of Dhp1 resulted in a generalized increase in Ser2-P and Tyr1-P signal, which was most noticeable at the 3' end of genes. Analysis of Ser2-P/total RNAPII ratios at the 3' end of genes suggests that the levels of Ser2 phosphorylation does not truly increase in Dhp1-deficient cells, but that the CTD remains phosphorylated on Ser2 for an extended period of time after cleavage (Fig. S5B). In contrast, Tyr1-P/total RNAPII ratios showed an increase in the maximal signal of Tyr1 phosphorylation in the *dhp1* mutant (Fig. S5B), suggesting that more CTD repeats have a phosphorylation mark on Tyr1 in Dhp1-depleted cells compared to wild-type cells. Prolonged Pcf11 binding at the 3' end of coding (Fig. S5C-F) and ncRNA (Fig. S5J-S5K) genes further support the functional relevance of sustained Ser2-P in Dhp1-deficient cells. The changes in CTD phosphorylation at Ser2 and Tyr1 in the *dhp1* mutant do not appear to be the result of reduced recruitment of the Dis2/Glc7 phosphatase, which showed similar binding between wild-type and Dhp1-deficient cells after normalization to total RNAPII (Fig. S5G-S5I). In contrast to Ser2-P and Tyr1-P, Ser5-P and Ser7-P signals were only slightly affected by the depletion of Dhp1 (Fig. S5L-S5M). We conclude that altered CTD phosphorylation is a conserved consequence of Dhp1/Rat1 inactivation.

2.2.4.7. The endonucleolytic activity of Ysh1 is essential for transcription termination at snoRNA genes.

The global transcription termination defects at mRNA and snoRNA genes in Dhp1-deficient cells suggest that coding and ncRNA genes are terminated by torpedo in fission yeast. Accordingly, this mechanism entails an endonucleolytic cleavage prior to

Dhp1-dependent transcription termination. To test whether the endoribonuclease subunit of the cleavage and polyadenylation specificity factor (CPSF) complex, Ysh1 (CPSF-73 in humans), is required for transcription termination of ncRNA genes in fission yeast, we analyzed RNAPII density in cells where Ysh1 was depleted from the nucleus using an anchor-away strategy. As expected, inactivation of Ysh1 resulted in a clear read-through phenotype at a protein-coding gene (Fig. 14A). Globally, the RNAPII 3' peak showed a clear shift downstream in the *ysh1* mutant as compared to control cells (Fig. 14D). Importantly, the consequences of Ysh1 nuclear depletion were comparable at the *snR99* gene (Fig. 14B), causing a marked downstream shift in the average RNAPII profile at snoRNA genes (Fig. 14E). The transcription termination defects observed at snoRNA genes in the *ysh1* mutant are consistent with the detection of read-through transcripts in this mutant (Fig. 14F). Interestingly, whereas Dhp1 was required for efficient transcription termination of both snoRNA and snRNA genes (Fig. 13B-13C and Fig. S5A), the nuclear depletion of Ysh1 did not result in read-through transcription at snRNA genes (Fig. 14C and Fig. S6A).

We next addressed whether the endonucleolytic activity of Ysh1 was required to promote termination of snoRNA transcription. To test this, we generated *ysh1* alleles expressing single amino acid substitutions at conserved histidine residues that were shown to contact critical zinc atoms in the active site of human CPSF-73 [42]. We used a "complementation after nuclear depletion" approach [43] by chromosomal integration of the different *ysh1* alleles as single copies into the *ysh1* anchor-away strain and confirmed that the corresponding proteins were expressed (Fig. S6B). After nuclear depletion of endogenous Ysh1 by rapamycin, the expression of catalytic mutant versions of Ysh1 resulted in growth arrest (Fig. S6C), consistent with the catalytic activity of Ysh1 being required for cell viability [44].

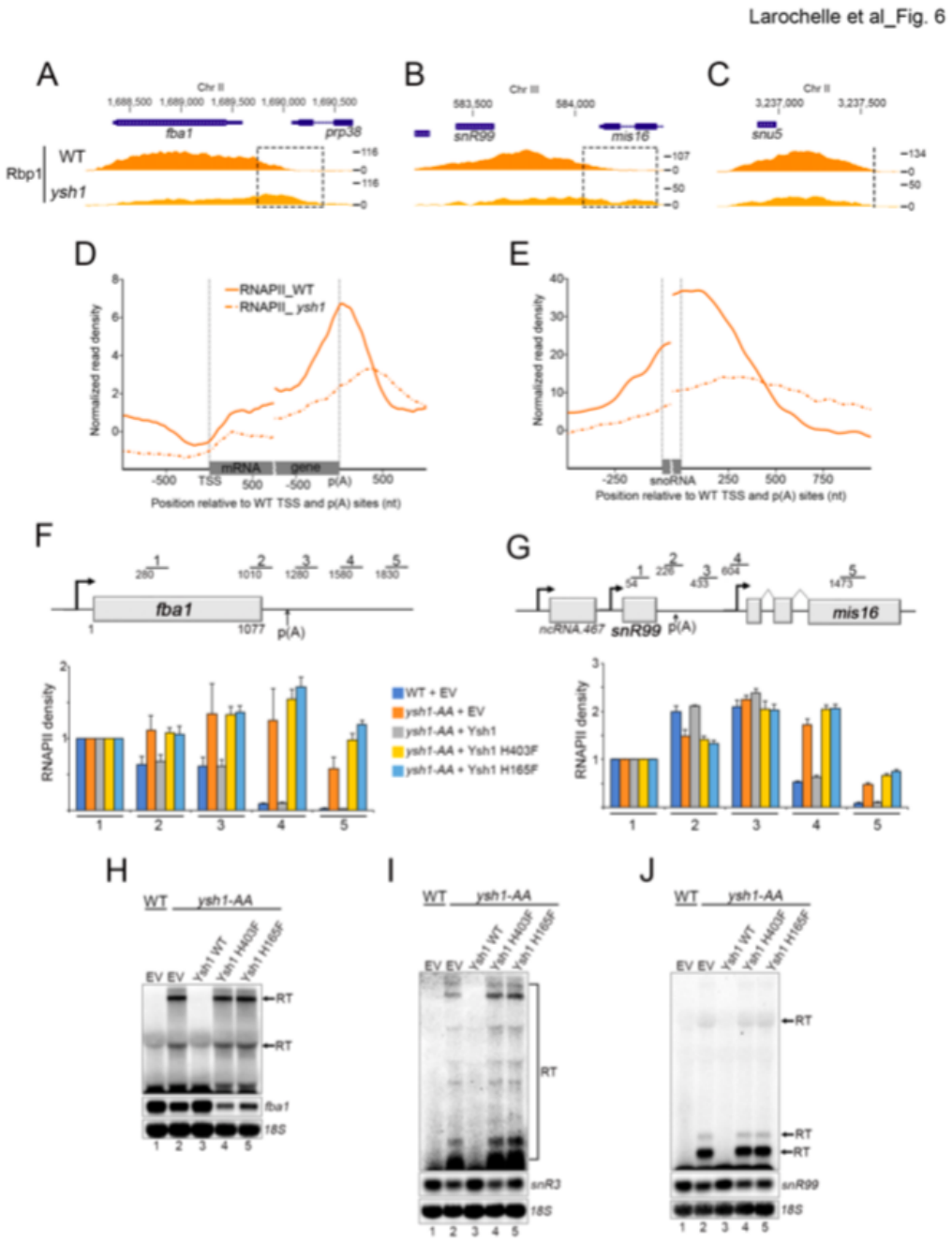


Figure 14. The endonucleolytic activity of Ysh1 is necessary for termination of snoRNA transcription.

(A-C) Normalized ChIP-seq signal of total RNAPII (Rbp1) in WT (top) and *ysh1* mutant (bottom) cells across the *fba1* mRNA (A), the *snR99* snoRNA (B), and the *snu5* snRNA (C) genes in rapamycin-treated minimal medium. The dashed-line rectangles highlight transcriptional read-through at *fba1* and *snR99* genes in Ysh1-deficient cells. (D-E)

Average ChIP-seq profile of RNAPII (Rpb1) in WT (solid lines) and Ysh1-deficient (dotted lines) cells across 4,755 mRNA (D) and 24 monocistronic snoRNA (E) genes in rapamycin-treated rich medium. (F-G) RNAPII ChIP-qPCR analysis on the *fba1* (F) and *snR99* (G) genes using extracts prepared from either wild-type (WT) or *ysh1* anchor-away (*ysh1-AA*) strains containing genomically integrated constructs that express the indicated versions of FLAG-tagged Ysh1 (WT, H403F, and H165F) as well as an empty vector (EV) control. Bars above the *fba1* and *snR99* genes show the positions of PCR products used for ChIP-qPCR analyses. Cells were grown in the presence of rapamycin to deplete endogenous Ysh1 from the nucleus. ChIP signals (percent of input) were normalized to region 1. Error bars indicate SD. *n* = 3 biological replicates from independent cultures. (H-J) Northern blot analysis of *fba1* (H), *snR3* (I), and *snR99* (J) genes using total RNA prepared from using total RNA prepared from the same WT (lane 1) and *ysh1* anchor-away (*ysh1-AA*, lanes 2-5) strains as in panels F-G. The position of read-through (RT) transcripts is indicated on the right. The 18S rRNA was used as a loading control.

We next used RNAPII ChIP-qPCR assays to examine the extent to which the mutant versions of Ysh1 restored the transcription termination defects induced by the nuclear depletion of endogenous Ysh1. As a control, expression of wild-type Ysh1 in the *ysh1* anchor-away strain prevented the rapamycin-dependent increase in RNAPII levels downstream of both protein- and snoRNA-coding genes (Fig. 14F-14G, compare dark blue, grey, and orange bars). In contrast, the endonuclease mutant versions of Ysh1 showed increase in RNAPII density at the 3' end of both mRNA and snoRNA genes (Fig. 14F-14G, compare yellow and light blue bars to grey bars). Consistently with these RNAPII ChIP-qPCR data, RNA analysis showed the accumulation of read-through transcripts in cells that expressed mutant versions of Ysh1 (Fig. 14H-14J, compare lanes 4-5 to lane 3; Fig. S6D). We conclude that the endonucleolytic activity of Ysh1 is required for 3' end processing and transcription termination of both mRNA and snoRNA genes.

2.2.5. Discussion

The existence of distinct pathways that promote termination of RNAPII transcription in the budding yeast *S. cerevisiae* has been well documented. In this organism, transcription of mRNA-coding genes is terminated by cleavage/polyadenylation factors (CPF), whereas the Nrd1-Nab3-Sen1 (NNS)-dependent pathway primarily terminates ncRNA transcription. Despite a fairly good understanding of how the NNS complex promotes termination at ncRNA genes, the conservation of NNS-like transcription termination across eukaryotic species has remained elusive. Unexpectedly, despite the presence of homologs for all of the NNS components, our results indicate that NNS-mediated termination is not conserved in the distantly related yeast *S. pombe*. In the absence of an NNS-like complex, our findings reveal that a universal cleavage-dependent mechanism that involves the conserved Dhp1 5'-3' torpedo nuclease is used to terminate transcription of both mRNA and ncRNA genes.

Two key indications support the absence of an NNS-like transcription termination pathway in fission yeast. First, Seb1 (Nrd1 homolog), Nab3, Sen1, and Dbl8 do not to form a stable complex in fission yeast [26,27,45]. Second, transcription termination defects were not detected in *nab3Δ*, *sen1Δ*, *dbl8Δ*, and *sen1Δ dbl8Δ* mutants (Fig. 9A-9B). These data suggest that while NNS components have been conserved between budding and fission yeasts, they have acquired species-specific functions. For instance, *S. pombe* Seb1 recognizes Ser2-phosphorylated RNAPII and is recruited to the 3' end of both mRNA and ncRNA genes (Fig. 9) [26,27]. In contrast, the Seb1 homolog in *S. cerevisiae*, Nrd1, preferentially binds to Ser5-phosphorylated RNAPII and is enriched at noncoding transcription units [2]. Also consistent with the idea that NNS components have functionally diverged between the evolution of budding and fission yeasts from a common ancestor, Sen1 is primarily associated with RNA polymerase III (RNAPIII)-transcribed genes in *S. pombe* where it antagonizes RNAPIII transcription

[45]. An NNS-like transcription termination pathway is also unlikely to exist in metazoans, as clear homologs or functional analogs of Nrd1 and Nab3 components have not been described. Conversely, a homolog for the yeast Sen1 helicase has been characterized in humans (Senataxin, SETX), and a role for SETX in promoting efficient transcription termination at model protein-coding genes has been described [46,47]. However, how human SETX globally contributes to transcription termination of mRNA genes remains to be demonstrated, as data presented in this study (Fig. 9A) as well as in *S. cerevisiae* [48] indicate that yeast Sen1 is not required for efficient transcription termination at protein-coding genes.

The finding that fission yeast does not have a functional NNS-like transcription termination pathway raised the intriguing question as to the nature of the mechanism used to promote termination at ncRNA genes. By combining genome-wide studies and functional analyses of the fission yeast mRNA 3' end processing machinery, we provide evidence supporting that a general cleavage-dependent, torpedo-mediated mechanism is used to promote transcription termination of protein-coding and ncRNA genes. Several observations support this conclusion: (i) mRNA 3' end processing (Rna14, Pcf11, and Ysh1) and termination (Dhp1) factors are specifically recruited at the 3' end of independently-transcribed snoRNA and snRNA genes (Fig. 9); (ii) mRNA 3' end processing factors are required for snoRNA synthesis (Fig. 10); (iii) polyadenylated pre-snoRNAs are produced from most snoRNA genes (Fig. 11); (iv) 3' end *cis*-sequences of protein-coding genes can promote snoRNA 3' end maturation and accumulation (Fig. 11); and (v) transcription termination at ncRNA genes requires endonucleolytic cleavage and the torpedo nuclease Dhp1 (Fig. 13-14). These data support a model in which cleavage and polyadenylation are essential steps for 3' end processing and transcription termination at both mRNA (Fig. 15A) and snoRNA (Fig. 15B) genes. Although it has been known that particular components of the mRNA 3' end processing machinery, such as Pcf11 and Glc7, are important for termination of short ncRNA genes in *S. cerevisiae* [49-51], these factors appear to function separately from the

cleavage and polyadenylation complex to contribute to NNS-dependent termination. This contrasts our findings using fission yeast, where the prevailing mechanism of transcription termination at ncRNA genes involves the cleavage and polyadenylation complex. Given the unexpected similarity of cleavage-dependent transcription termination at mRNA and snoRNA genes, it will be interesting to understand the mechanism used to distinguish between mRNAs and snoRNAs; whereas the mature mRNA needs to retain its poly(A) tail (Fig. 15A), the polyadenylated 3'-extended pre-snoRNA is targeted by Pab2-dependent exosome-mediated 3' end trimming to yield a mature non-polyadenylated snoRNA (Fig. 15B). We speculate that co-transcriptional decoration of nascent pre-mRNAs with nuclear export factors prevents nuclear retention and Pab2-dependent 3' end trimming (Fig. 15A), whereas the nucleolar-targeting signals of snoRNA ribonucleoproteins (snoRNPs) assembled during snoRNA transcription [52] would promote nuclear retention and 3' end maturation (Fig. 15B).

As for snoRNA genes, termination of snRNA transcription in *S. cerevisiae* appears primarily dependent on NNS [53]. Surprisingly, we found that whereas transcription termination at both snoRNA and snRNA genes required Dhp1 in *S. pombe* (Fig. 13 and Fig. S5A), termination of snRNA transcription was not affected in Ysh1-deficient cells (Fig. 14 and Fig. S6A). This result suggests that the 5' entry point for Dhp1 loading at snRNA genes may be mediated independently of the nuclease activity of the mRNA 3' end processing machinery. Interestingly, read-through transcription at the *S. pombe* U2 snRNA gene was previously reported in a temperature sensitive mutant of *pac1* [54], which encodes a homolog of the *S. cerevisiae* RNase III-like endonuclease Rnt1. Rnt1 cleavage was in fact shown to function in a failsafe mechanism of snRNA transcription termination in *S. cerevisiae* that relies on Rat1 5'-3' exonucleolytic activity [55,56]. Although the identity of the nuclease responsible for cleavage at *S. pombe* snRNA genes remains to be determined, our data suggest that snRNA transcription termination in fission yeast is more closely related to humans than to *S. cerevisiae*, as mRNA 3'

end processing factors [57] and the torpedo nuclease XRN2 [6] are important for efficient transcription termination of human snRNA genes.

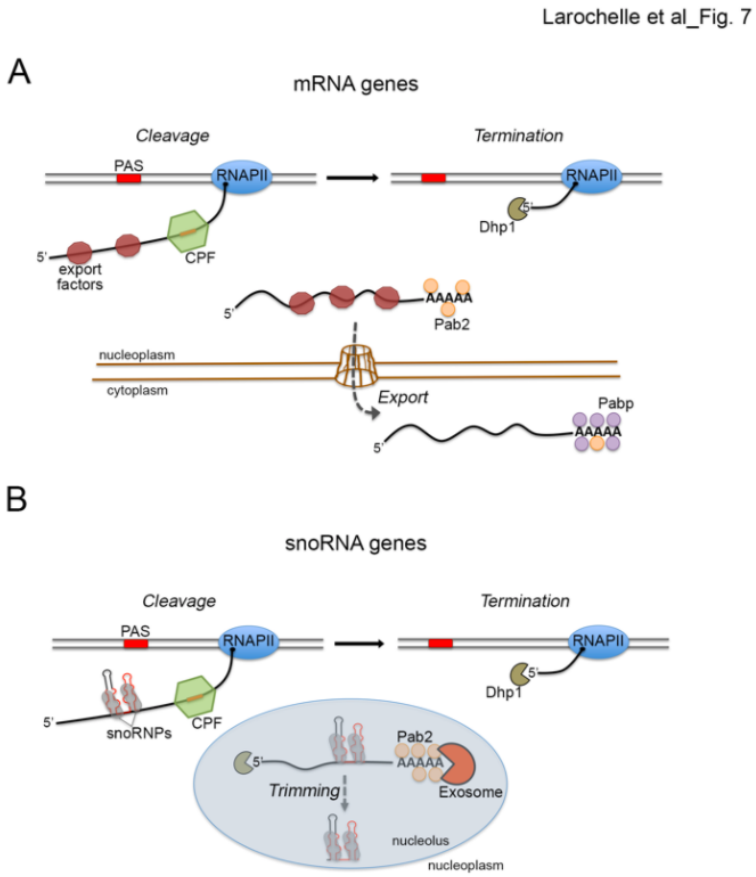


Figure 15. Model for 3' end processing and transcription termination of mRNA and snoRNA genes in fission yeast.

(A-B) Recruitment of cleavage and polyadenylation factors (CPF) by poly(A) signals is a common feature of mRNA and snoRNA genes (see *Cleavage*). Endonucleolytic cleavage by the CPF-associated Ysh1 nuclease will generate an RNAPII-bound unprotected 5' end that will be targeted by the 5'-3' exonuclease Dhp1, contributing to termination of mRNA and snoRNA transcription (see *Termination*). The co-transcriptional recruitment of specific export factors to nascent mRNAs (A) may represent a decisive step that prevents Pab2-dependent exosome-mediated RNA processing in the nucleus (B).

Our studies also elicit interesting questions as to the functional significance of RNAPII CTD tyrosine phosphorylation (Tyr1-P) in fission yeast. Our data indicated that the pattern of Tyr1-P on the CTD of *S. pombe* RNAPII at the 3' end of genes (Fig. 12) differed from the distribution of Tyr1-P in *S. cerevisiae* [22]. Accordingly, although Tyr1-P levels increase along coding regions in both budding and fission yeasts, Tyr1-P signal decreases upstream of the poly(A) site in *S. cerevisiae*, whereas Tyr1-P signal persisted downstream of the poly(A) site in *S. pombe*. This differing pattern of Tyr1-P around the poly(A) was unexpected, as Tyr1-P in *S. cerevisiae* was proposed to inhibit binding of CID-containing termination factors (Pcf11 and Rtt103) to Ser2-P CTD and prevent premature termination in coding regions [22], a role that would also be anticipated to contribute to general transcription in *S. pombe*. Yet, the similar distribution of Ser2-P and Tyr1-P profiles downstream of the poly(A) site in *S. pombe* (Fig. 12) argues for diverging roles of Tyr1-P between budding and fission yeasts. Consistent with this idea, a truncated version of the CTD in which a phenylalanine replaces Tyr1 (Y1F) in each heptad repeat results in lethality in *S. cerevisiae* [58], whereas the analogous truncated Y1F CTD mutant is viable in *S. pombe* [59]. Alternatively, Tyr1-P could also function to impair binding of CID-containing termination factors in *S. pombe* (Seb1, Pcf11, and Rhn1) by coordinating their recruitment to Rpb1 downstream of the poly(A) in a CTD repeat-specific manner.

Our genome-wide analysis of RNAPII occupancy in Dhp1-deficient cells revealed that a 5'-3' "torpedo" nuclease is necessary for transcription termination of most protein-coding genes in fission yeast, consistent with findings using *S. cerevisiae* and human cells [6,8,32]. In addition to transcription termination defects, elevated levels of Ser2-P signal were found at model protein-coding genes in a temperature-sensitive mutant of *S. cerevisiae rat1* [40]. Our data extend these finding by showing that (i) this observation is conserved in fission yeast, (ii) that Dhp1 depletion also causes elevated levels of Tyr1 phosphorylation, and (iii) that Dhp1 deficiency affects Tyr1 and Ser2 CTD phosphorylation at the genome-wide level. Although further studies will be needed to

investigate how Dhp1/Rat1 nucleases modulate CTD phosphorylation at the 3' end of genes, which may involve competition with the recruitment of CTD kinases [60], this finding highlights the important functional relationship between CTD phosphorylation and transcription termination.

Given that NNS-like transcription termination does not appear to be conserved in mammalian cells, it remains unclear how independently transcribed human snoRNAs, such as U3, U8, U13, and the human telomerase RNA, acquire their mature 3' end. The fact that we have recently uncovered a polyadenylation-dependent 3' end maturation pathway for the human telomerase RNA that depends on the nuclear poly(A)-binding protein PABPN1 [61] supports the existence of an evolutionarily conserved role for Pab2/PABPN1 in the maturation of independently transcribed snoRNAs. Accordingly, it will be interesting to determine whether 3' end processing and transcription termination of human snoRNA genes that are characterized by intergenic location and independent transcription unit depend on the mRNA cleavage and polyadenylation complex, as unveiled by our data using fission yeast.

2.2.6. Experimental procedures

2.2.6.1. Yeast strains and media

A list of all *S. pombe* strains used in this study is provided in Table S4. Fission yeast cells were grown at 30°C in yeast extract medium with adenine, uracil and amino acid supplements (YES) or in Edinburgh minimal media (EMM) supplemented with adenine, uracil and appropriate amino acids. Additional details regarding strains and growth conditions are provided in Supplemental Experimental Procedures.

2.2.6.2. Chromatin immunoprecipitation (ChIP) assays

ChIP-qPCR and ChIP-seq experiments were performed as described previously [26]. The antibodies used are described in Supplemental Experimental Procedures.

2.2.6.3. RNA analyses

Preparation of total fission yeast RNA, Northern blotting, and RNase H cleavage assays were performed as described previously [30,31] and described in the Supplemental Experimental Procedures.

2.2.6.4. Computational Methods

Detailed data analysis of ChIP-seq (Table S1) and 3'READS [36] experiments are described in the Supplemental Experimental Procedures, including a link to visualize processed data.

2.2.6.5. Accession number

All raw and processed data have been deposited in the GEO database under accession number GSE115595.

2.2.7. Acknowledgments

We thank Michelle Scott, and François Robert for critical reading of the manuscript; Luc Gaudreau for *S. cerevisiae* strains; the sequencing platforms of the McGill University and Génome Québec Innovation Centre; Calcul Québec and Compute Canada provided the computing infrastructure used to analyze the data. This work was supported by funding from the Natural Sciences and Engineering Research Council of Canada (NSREC) to F.B. (RGPIN-2017-05482) and P-É. J. (435710-2013), and by funding from the National Institute of General Medical Sciences to B.T. (GM084089); P-É.J. is supported by the Fonds de recherche du Québec – Santé (FRQS) and F.B. holds a Canada Research Chair in Quality Control of Gene Expression.

2.2.8. Author contributions

M.L. and F.B. conceived the study and the experimental frame, while M-A.R and P-E.J. planned and designed the pipelines for processing the genome-wide data. M.L. prepared chromatin extracts for ChIPs, performed ChIP-seq, including QCs and library preparations. M.L. performed most of the RNA analyses with help from J-N.H.. J-N.H. made the Ysh1 constructs and performed the phenotypic characterization of the ysh1 and rna14 anchor-away mutants. X.L. prepared the 3'READS libraries with help from B.T.. D.M. and S.R. helped with the preparation of the ChIP-seq libraries. M-A.R. and P-E.J. performed all of the bioinformatics analyses of the ChIP-seq and 3'READS data with the help X.L.. M.L., M-A.R, P-E.J., and F.B. prepared and finalized the figures. F.B. wrote the manuscript with help of P-E.J., which was reviewed by all authors

2.2.9. References

1 Proudfoot, N. J. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science* 352, aad9926, doi:10.1126/science.aad9926 (2016).

2 Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat Rev Mol Cell Biol* 16, 190-202, doi:10.1038/nrm3943 (2015).

3 Casanal, A. *et al.* Architecture of eukaryotic mRNA 3'-end processing machinery. *Science*, doi:10.1126/science.aao6535 (2017).

4 Shi, Y. & Manley, J. L. The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes Dev* 29, 889-897, doi:10.1101/gad.261974.115 (2015).

5 Baejen, C. *et al.* Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes. *Mol Cell* 66, 38-49 e36, doi:10.1016/j.molcel.2017.02.009 (2017).

6 Fong, N. *et al.* Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Mol Cell* 60, 256-267, doi:10.1016/j.molcel.2015.09.026 (2015).

7 Kim, M. *et al.* The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* 432, 517-522, doi:10.1038/nature03041 (2004).

8 West, S., Gromak, N. & Proudfoot, N. J. Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* 432, 522-525, doi:10.1038/nature03035 (2004).

9 Gudipati, R. K., Villa, T., Boulay, J. & Libri, D. Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. *Nat Struct Mol Biol* 15, 786-794, doi:10.1038/nsmb.1460 (2008).

10 Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S. & Meinhart, A. The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nat Struct Mol Biol* 15, 795-804 (2008).

11 Carroll, K. L., Pradhan, D. A., Granek, J. A., Clarke, N. D. & Corden, J. L. Identification of cis elements directing termination of yeast nonpolyadenylated snoRNA transcripts. *Mol Cell Biol* 24, 6241-6252 (2004).

12 Creamer, T. J. *et al.* Transcriptome-wide binding sites for components of the *Saccharomyces cerevisiae* non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS genetics* 7, e1002329, doi:10.1371/journal.pgen.1002329 (2011).

13 Porrua, O. & Libri, D. A bacterial-like mechanism for transcription termination by the Sen1p helicase in budding yeast. *Nat Struct Mol Biol* 20, 884-891, doi:10.1038/nsmb.2592 (2013).

14 Tudek, A. *et al.* Molecular basis for coordinating transcription termination with noncoding RNA degradation. *Mol Cell* 55, 467-481, doi:10.1016/j.molcel.2014.05.031 (2014).

15 Vasiljeva, L. & Buratowski, S. Nrd1 interacts with the nuclear exosome for 3' processing of RNA polymerase II transcripts. *Mol Cell* 21, 239-248 (2006).

16 Zaborowska, J., Egloff, S. & Murphy, S. The pol II CTD: new twists in the tail. *Nat Struct Mol Biol* 23, 771-777, doi:10.1038/nsmb.3285 (2016).

17 Suh, H. *et al.* Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II. *Mol Cell* 61, 297-304, doi:10.1016/j.molcel.2015.12.021 (2016).

18 Kim, H. *et al.* Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nat Struct Mol Biol* 17, 1279-1286, doi:10.1038/nsmb.1913 (2010).

19 Mayer, A. *et al.* Uniform transitions of the general RNA polymerase II transcription complex. *Nat Struct Mol Biol* 17, 1272-1278, doi:10.1038/nsmb.1903 (2010).

20 Lunde, B. M. *et al.* Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. *Nat Struct Mol Biol* 17, 1195-1201, doi:10.1038/nsmb.1893 (2010).

21 Meinhart, A. & Cramer, P. Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. *Nature* 430, 223-226, doi:10.1038/nature02679 (2004).

22 Mayer, A. *et al.* CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science* 336, 1723-1725, doi:10.1126/science.1219651 (2012).

23 Schreieck, A. *et al.* RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7. *Nat Struct Mol Biol* 21, 175-179, doi:10.1038/nsmb.2753 (2014).

24 Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* 489, 101-108, doi:10.1038/nature11233 (2012).

25 Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* 453, 1239-1243 (2008).

26 Lemay, J. F. *et al.* The Nrd1-like protein Seb1 coordinates cotranscriptional 3' end processing and polyadenylation site selection. *Genes Dev* 30, 1558-1572, doi:10.1101/gad.280222.116 (2016).

27 Wittmann, S. *et al.* The conserved protein Seb1 drives transcription termination by binding RNA polymerase II and nascent RNA. *Nature communications* 8, 14861, doi:10.1038/ncomms14861 (2017).

28 Banerjee, A., Apponi, L. H., Pavlath, G. K. & Corbett, A. H. PABPN1: molecular function and muscle disease. *The FEBS journal* 280, 4230-4250, doi:10.1111/febs.12294 (2013).

29 Larochelle, M., Lemay, J. F. & Bachand, F. The THO complex cooperates with the nuclear RNA surveillance machinery to control small nucleolar RNA expression. *Nucleic Acids Res* 40, 10240-10253, doi:10.1093/nar/gks838 (2012).

30 Lemay, J. F. *et al.* The nuclear poly(A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. *Mol Cell* 37, 34-45 (2010).

31 Lemay, J. F. *et al.* The RNA exosome promotes transcription termination of backtracked RNA polymerase II. *Nat Struct Mol Biol* 21, 919-926, doi:10.1038/nsmb.2893 (2014).

32 Kim, D. U. *et al.* Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nat Biotechnol* 28, 617-623, doi:10.1038/nbt.1628 (2010).

33 Larochelle, M., Hunyadkurti, J. & Bachand, F. Polyadenylation site selection: linking transcription and RNA processing via a conserved carboxy-terminal domain (CTD)-interacting protein. *Curr Genet* 63, 195-199, doi:10.1007/s00294-016-0645-8 (2017).

34 Ding, L., Laor, D., Weisman, R. & Forsburg, S. L. Rapid regulation of nuclear proteins by rapamycin-induced translocation in fission yeast. *Yeast* 31, 253-264, doi:10.1002/yea.3014 (2014).

35 Hoque, M. *et al.* Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat Methods* 10, 133-139, doi:10.1038/nmeth.2288 (2013).

36 Liu, X. *et al.* Comparative analysis of alternative polyadenylation in *S. cerevisiae* and *S. pombe*. *Genome Res* 27, 1685-1695, doi:10.1101/gr.222331.117 (2017).

37 Mata, J. Genome-wide mapping of polyadenylation sites in fission yeast reveals widespread alternative polyadenylation. *RNA Biol* 10 (2013).

38 Schlackow, M. *et al.* Genome-wide analysis of poly(A) site selection in *Schizosaccharomyces pombe*. *RNA* 19, 1617-1631, doi:10.1261/rna.040675.113 (2013).

39 Harlen, K. M. & Churchman, L. S. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat Rev Mol Cell Biol*, doi:10.1038/nrm.2017.10 (2017).

40 Jimeno-Gonzalez, S., Schmid, M., Malagon, F., Haaning, L. L. & Jensen, T. H. Rat1p maintains RNA polymerase II CTD phosphorylation balance. *RNA* 20, 551-558, doi:10.1261/rna.041129.113 (2014).

41 Orlando, D. A. *et al.* Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell reports* 9, 1163-1170, doi:10.1016/j.celrep.2014.10.018 (2014).

42 Mandel, C. R. *et al.* Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature* 444, 953-956 (2006).

43 Jeronimo, C. & Robert, F. Kin28 regulates the transient association of Mediator with core promoters. *Nat Struct Mol Biol* 21, 449-455, doi:10.1038/nsmb.2810 (2014).

44 Ryan, K., Calvo, O. & Manley, J. L. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. *RNA* 10, 565-573 (2004).

45 Legros, P., Malapert, A., Niinuma, S., Bernard, P. & Vanoosthuyse, V. RNA processing factors Swd2.2 and Sen1 antagonize RNA Pol III-dependent transcription and the localization of condensin at Pol III genes. *PLoS genetics* 10, e1004794, doi:10.1371/journal.pgen.1004794 (2014).

46 Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol Cell* 42, 794-805, doi:10.1016/j.molcel.2011.04.026 (2011).

47 Zhao, D. Y. *et al.* SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination. *Nature* 529, 48-53, doi:10.1038/nature16469 (2016).

48 Schaughency, P., Merran, J. & Corden, J. L. Genome-wide mapping of yeast RNA polymerase II termination. *PLoS genetics* 10, e1004632, doi:10.1371/journal.pgen.1004632 (2014).

49 Grzechnik, P., Gdula, M. R. & Proudfoot, N. J. Pcf11 orchestrates transcription termination pathways in yeast. *Genes Dev* 29, 849-861, doi:10.1101/gad.251470.114 (2015).

50 Kim, M. *et al.* Distinct pathways for snoRNA and mRNA termination. *Mol Cell* 24, 723-734 (2006).

51 Nedea, E. *et al.* The Glc7 phosphatase subunit of the cleavage and polyadenylation factor is essential for transcription termination on snoRNA genes. *Mol Cell* 29, 577-587 (2008).

52 Pradet-Balade, B. *et al.* CRM1 controls the composition of nucleoplasmic pre-snoRNA complexes to licence them for nucleolar transport. *EMBO J* 30, 2205-2218, doi:10.1038/emboj.2011.128 (2011).

53 Jamonnak, N. *et al.* Yeast Nrd1, Nab3, and Sen1 transcriptome-wide binding maps suggest multiple roles in post-transcriptional RNA processing. *RNA* 17, 2011-2025, doi:10.1261/rna.2840711 (2011).

54 Nabavi, S. & Nazar, R. N. Cleavage-induced termination in U2 snRNA gene expression. *Biochem Biophys Res Commun* 393, 461-465, doi:10.1016/j.bbrc.2010.02.023 (2010).

55 Ghazal, G. *et al.* Yeast RNase III triggers polyadenylation-independent transcription termination. *Mol Cell* 36, 99-109, doi:10.1016/j.molcel.2009.07.029 (2009).

56 Rondon, A. G., Mischo, H. E., Kawauchi, J. & Proudfoot, N. J. Fail-safe transcriptional termination for protein-coding genes in *S. cerevisiae*. *Mol Cell* 36, 88-98, doi:10.1016/j.molcel.2009.07.028 (2009).

57 O'Reilly, D. *et al.* Human snRNA genes use polyadenylation factors to promote efficient transcription termination. *Nucleic Acids Res* 42, 264-275, doi:10.1093/nar/gkt892 (2014).

58 West, M. L. & Corden, J. L. Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. *Genetics* 140, 1223-1233 (1995).

59 Schwer, B. & Shuman, S. Deciphering the RNA polymerase II CTD code in fission yeast. *Mol Cell* 43, 311-318, doi:10.1016/j.molcel.2011.05.024 (2011).

60 Jimeno-Gonzalez, S., Haaning, L. L., Malagon, F. & Jensen, T. H. The yeast 5'-3' exonuclease Rat1p functions during transcription elongation by RNA polymerase II. *Mol Cell* 37, 580-587, doi:10.1016/j.molcel.2010.01.019 (2010).

61 Nguyen, D. *et al.* A Polyadenylation-Dependent 3' End Maturation Pathway Is Required for the Synthesis of the Human Telomerase RNA. *Cell reports* 13, 2244-2257, doi:10.1016/j.celrep.2015.11.003 (2015).

2.2.10. Supplemental information

2.2.10.1. Supplemental information inventory

Supplemental Information includes six figures, four tables, Supplemental, Experimental Procedures, and Supplemental References.

SUPPLEMENTARY FIGURES :

- **Figure S1, related to Figure 1.** *S. pombe* mRNA 3' end processing and transcription termination factors are not recruited to intronic snoRNA genes.
- **Figure S2, related to Figure 2.** *S. pombe* Ysh1 and Rna14 are essential for viability and mRNA synthesis.
- **Figure S3, related to Figures 3.** Effects of Pab2, Seb1, and Pcf11 deficiencies on mRNA polyadenylation.
- **Figure S4, related to Figure 4.** Genowe-wide correlation between ChIP-seq experiments.
- **Figure S5, related to Figure 5.** Dhp1 influences the pattern of Ser2 and Tyr1 CTD phosphorylation at the 3' end of genes.
- **Figure S6, related to Figure 6.** The endonucleolytic activity of Ysh1 is required for snoRNA synthesis.

SUPPLEMENTARY TABLES:

Is not include here.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

SUPPLEMENTAL REFERENCES

2.2.10.2. Supplementary figures

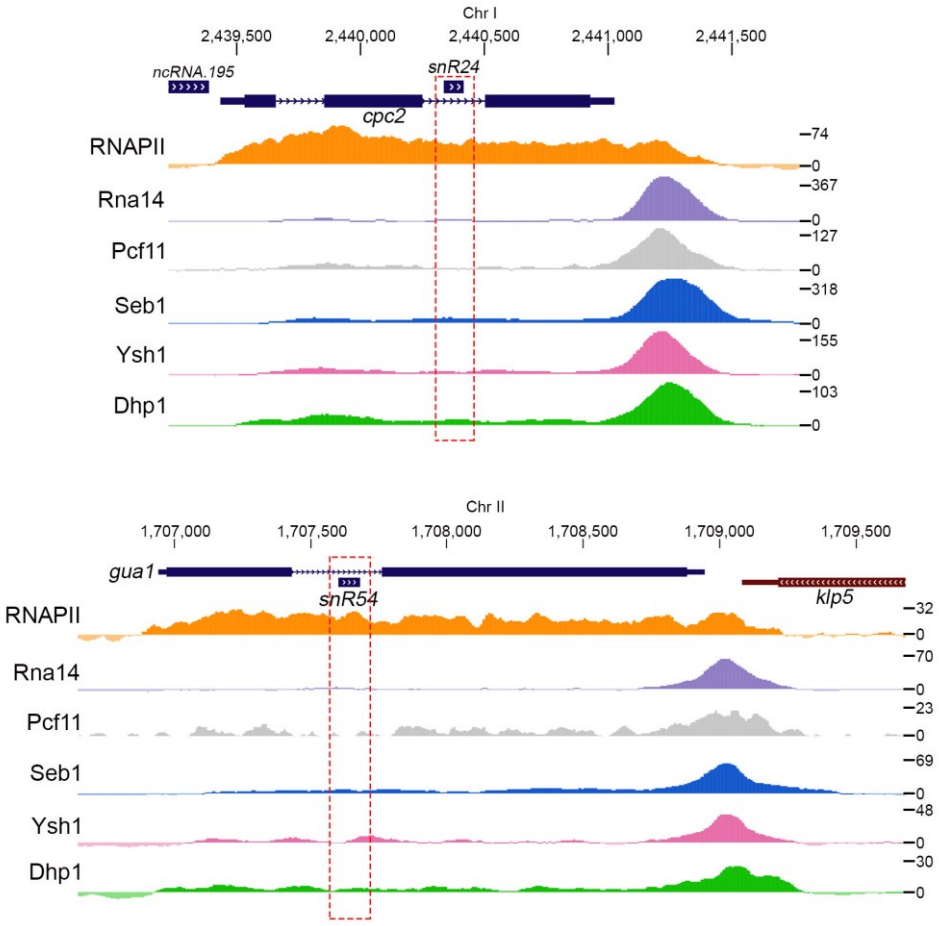


Figure S1. related to Figure 9. *S. pombe* mRNA 3' end processing and transcription termination factors are not recruited to intronic snoRNA genes. Normalized ChIP-seq signal of total RNAPII as well as the indicated mRNA 3' end processing and transcription termination factors across the *cpc2* (top) and *gua1* (bottom) mRNA genes. Dashed-line red rectangles show the absence of Rna14, Pcf11, Seb1, Ysh1, and Dhp1 enrichment at intronic *snR24* (top) and *snR54* (bottom) snoRNA genes.

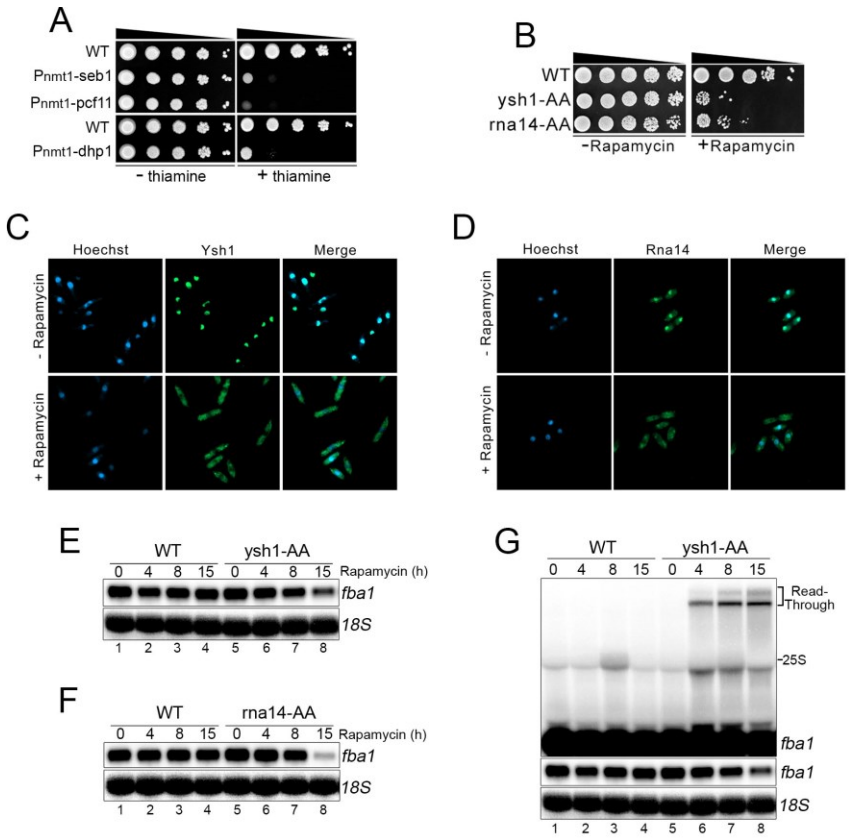


Figure S2. related to Figure 10. *S. pombe* Ysh1 and Rna14 are essential for viability and mRNA synthesis.

(A) Ten-fold serial dilutions of wild-type (WT), P_{nmt1}*-seb1,* P_{nmt1}*-pcf11,* and P_{nmt1}*-dhp1* cells were spotted on thiamine-free (left) or thiamine-containing (right) minimal media. (B) Ten-fold serial dilutions of wild-type (WT), *ysh1* anchor-away *(ysh1-AA),* and *rna14-AA* cells were spotted on rapamycin-free (left) or rapamycin-containing (right) minimal media. (C-D). Representative pictures of Ysh1-FRB-GFP (C) and Rna14-FRB-GFP (D) re-localization from the nucleus to the cytoplasm 3h after rapamycin treatment (bottom panels), whereas nuclear localization was observed in the absence of Rapamycin (top panels). (E-F) Rapamycin-dependent inhibition of mRNA synthesis in *ysh1-AA* (E) and *rna14-AA* (F) strains. Northern blot analysis of total RNA prepared from WT and the indicated anchor-away mutants at 0h, 4h, 8h, and 15h after treatment with rapamycin. The blot was hybridized with an antisense RNA probe complementary to the *fba1* mRNA. (G) Rapamycin-dependent accumulation of *fba1* read-through transcripts in the *ysh1-AA* strain. Note the accumulation of *fba1* read-through transcripts 4h after the addition of rapamycin in the *ysh1-AA* strain (lanes 5-8), but not in the wild-type (WT) strain (lanes 1-4). (E-G) The 18S rRNA was used as a loading control.

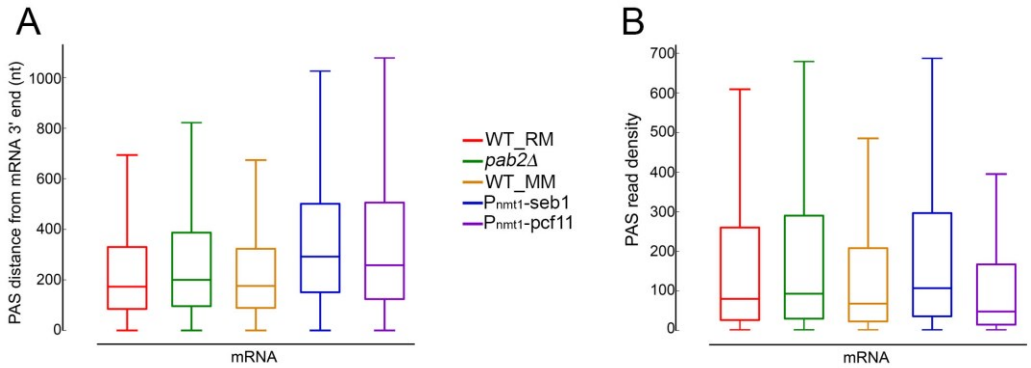


Figure S3. related to Figure 11. Effects of Pab2, Seb1, and Pcf11 deficiencies on mRNA polyadenylation.

(A-B) Distribution of the distance calculated between the strongest poly(A) site (PAS) and the mRNA stop codon as determined by 3'READS (A) and the sum of the read density for all of the poly(A) sites associated to a mRNA in each condition (B).

ChIP target	*Pearson Correlation Coefficients*		
Ser5-P B1	1	0.961	
Ser5-P B2	0.961	1	
Ser7-P B1	1	0.9698	
Ser7-P B2	0.9698	1	
Ser2-P B1	1	0.9613	
Ser2-P B2	0.9613	1	
Tyr1-P B1	1	0.9572	
Tyr1-P B2	0.9572	1	
Rpb1 B1	1	0.941	0.9276
Rpb1 B2	0.941	1	0.9614
Rpb1 B3	0.9276	0.9614	1
Rpb3 B1	1	0.9329	
Rpb3 B3	0.9329	1	

Figure S4. related to Figure 12. Genowe-wide correlation between ChIP-seq experiments.

(C) Genome-wide Pearson correlation coefficients between independent ChIP-seq experiments using antibodies targeting the indicated proteins or CTD modifications (B1, batch 1; B2, batch 2; B3, batch 3).

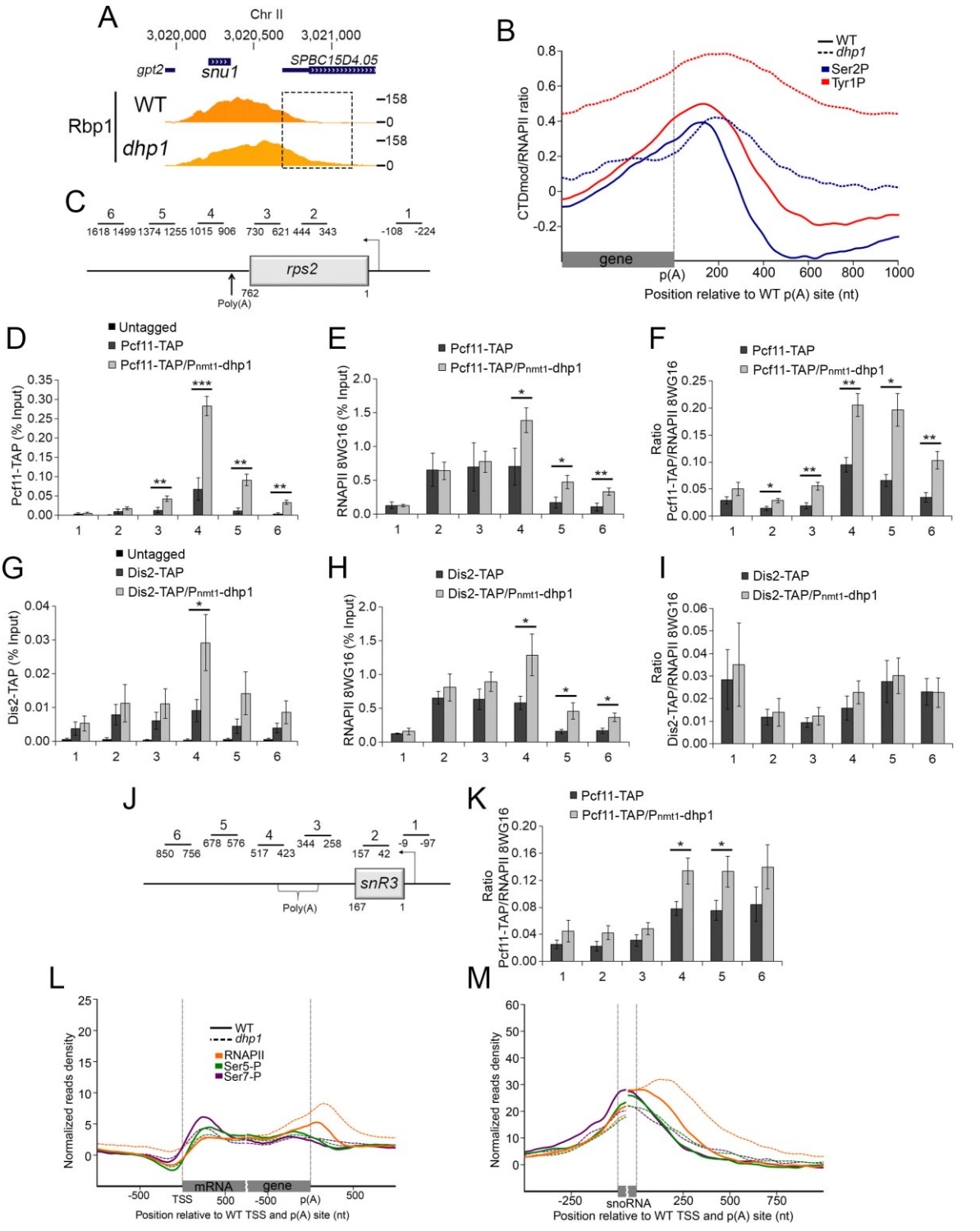


Figure S5. related to Figure 13. Dhp1 influences the pattern of Ser2 and Tyr1 CTD phosphorylation at the 3' end of genes.

(A) Normalized ChIP-seq signal of RNAPII subunit Rbp1 across the *snu1* snRNA gene in WT (top) and Dph1-depleted (bottom) strains. (B) Average ChIP-seq profile of Ser2-P (blue) and Tyr1-P (red) normalized by the RNAPII signal in a WT (solid lines) and Dhp1-depleted strain (dotted lines) centered on the polyA sites (p(A)) of 4,755 mRNA and 24 snoRNA genes with a PAS in minimal medium. (C) Bars above the *rps2 gene* show the positions of PCR products used for ChIP-qPCR analyses in panels D-I. (D) ChIP assays using a TAP-tagged version of Pcf11 or an untagged control strain at the *rps2* gene in wild-type and P_{nmt1}-*dhp1* strains in the presence of thiamine to deplete Dhp1. Input and copurified DNA were quantified by qPCR using primers shown in panel C. (E) ChIP assays performed using an RNAPII-specific antibody at the *rps2* gene in wild-type and P_{nmt1}-*dhp1* strains in the presence of thiamine to deplete Dhp1. (F) Density of Pcf11 relative to RNAPII at the *rps2* gene in wild-type and P_{nmt1}-*dhp1* strains in the presence of thiamine. Data and error bars represent the average and standard deviation from three biological replicates. (G-I) Same as for D-F but using a TAP-tagged version of Dis2 instead of Pcf11. (J) Bars above the *snR3 snoRNA gene* show the positions of PCR products used for ChIP-qPCR analyses in panel K. (K) Density of Pcf11 relative to RNAPII at the *snR3* gene in wild-type and P_{nmt1}-*dhp1* strains in the presence of thiamine. Data and error bars represent the average and standard deviation from three biological replicates. *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$ (Student's t-test). (L-M) Average ChIP-seq profile of total RNAPII (Rpb1, orange), Ser5-P (green), Ser7-P (purple) in WT (solid lines) and Dhp1-depleted (dotted lines) strain across 4,755 mRNA (L) and across 24 snoRNA (M) with a PAS identified in minimal medium.

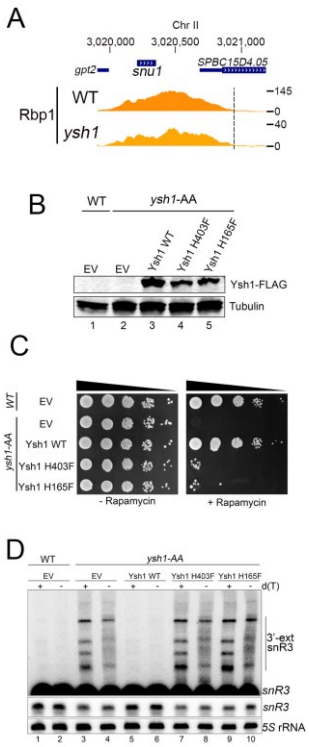


Figure S6., related to Figure 14. The endonucleolytic activity of Ysh1 is required for snoRNA synthesis.

(A) Normalized ChIP-seq signal of RNAPII subunit Rbp1 across the *snu1* snRNA gene in WT (top) and Ysh1-depleted (bottom) strains. (B) Western blot analysis of extracts prepared from WT cells (lane 1) as well as from *ysh1-AA* cells (lanes 2-5) in which an empty vector (EV) or constructs that express FLAG-tagged versions of WT Ysh1 (lane 3), H403F (lane 4) and H165F (lane 5) mutant versions of Ysh1. Cells were grown in the presence of rapamycin to deplete endogenous Ysh1. Tubulin was used as a loading control. (C) Ten-fold serial dilutions of wild-type and *ysh1-AA* cells that were transformed with an empty vector (EV) or constructs that express wild-type (WT) and mutant versions of Ysh1 were spotted on rapamycin-free (left) or rapamycin-containing (right) media. (D) Total RNA prepared from wild-type (lanes 1-2) and *ysh1-AA* (lanes 3-10) strains was treated with RNase H in the presence of a DNA oligonucleotide complementary to H/ACA class snoRNA *snR3*. RNase H reactions were performed in the presence (+) or absence (−) of oligo(dT). 5S rRNA was used as a loading control. Expression of catalytically inactive versions of Ysh1 resulted in reduced levels of mature *snR3* and the accumulation of 3-extended polyadenylated *snR3* precursors (lanes 7-10) similar to *ysh1-AA* cells transformed with the empty vector (EV) control (lanes 3-4).

2.2.10.3. Supplementary methods

2.2.10.3.1. Yeast strains and media

A list of all *S. pombe* strains used in this study is provided in Table S4. Cells were routinely grown at 30°C in Edinburgh minimal media (EMM) or in yeast extract medium (YES) supplemented with adenine, uracil, histidine and leucine. Conditional strains, in which the genomic copy of the essential genes *pcf11, dhp1* and *seb1* are expressed from the thiamine-sensitive *nmt1* promoter (P_{nmt1}), were repressed by the addition of 60 µM of thiamine in the EMM media for 12-15 hours as previously described (Lemay *et al.* 2014; Lemay *et al.* 2016). Depletion of nuclear Ysh1 and Rna14 were done by an anchor-away strateby (Ding *et al.* 2014). Briefly, wild-type (WT), Ysh1-FRB-GFP (*ysh1-AA*) and Rna14-FRB-GFP (*rna14-AA*) strains were grown in either EMM without leucine or YES media, and rapamycin was added or not at a final concentration of 2.5 µg/ml for 4-15h. Cells were collected at OD_{600nm} of ~0.5-0.8. Gene disruptions and C-terminal tagging of proteins were performed by PCR-mediated gene targeting (Bahler *et al.* 1998) using lithium acetate method for yeast transformation. Expression of tagged proteins was confirmed by western blotting and knockouts strains by the absence of RNA by RT-PCR.

2.2.10.3.2. Growth assays

Exponential cultures of *S. pombe* cells were adjusted to an OD_{600nm} of 1.0, and serially diluted 10-fold using water. Each dilution was spotted (3 µl/spot) on EMM plates with or without 15 µM of thiamine in the case of strains using the thiamine-sensitive *nmt1*

promoter (P_{nmt1}), or on YES plates containing DMSO or 2.5 µg/ml rapamycin for the analysis of anchor-away mutants. Plates were incubated at 30°C for 2 to 7 days.

2.2.10.3.3. Microscopy

Ysh1-GFP and Rna14-GFP localization was detected by using fluorescence microscopy. Briefly, precultures grown in EMM supplemented with adenine, uracil and histidine (EMM Leu$^-$) were used to inoculate larger 25-ml cultures that we grown in EMM leu$^-$ to early log phase (OD_{600nm} 0.25-0.3). Rapamycin was then added to a final concentration of 2.5µg/ml (same volume of DMSO was added in parallel as control) and samples were taken at 0h, 0.5h, 1h, 2h, 3h, 4h, and 5h, which were diluted 1:10 in water. Nuclei were stained using Hoechst 33342 for 5 min (0.2mg/ml) and live cells were mounted on a 1.2 % agarose/EMM leu$^-$ pad as described previously (Waddle *et al.* 1996). GFP-tagged proteins and nuclei were detected at 470nm and 365nm, respectively, using a Colibri system (Carl Zeiss Canada, Toronto, ON, Canada) on a Zeiss Axio Observer Z1 inverted microscope with a 60x/1.4 oil objective. Data were analysed using the ZEN black software (Carl Zeiss Canada).

2.2.10.3.4. RNA preparation and analyses

Total RNA was extracted using the hot-acid phenol method. RNA samples were resolved on agarose-formaldehyde gel or on 6 % polyacrylamide-8M urea gels, transferred onto nylon membranes and probed as described previously (Lemay *et al.* 2016). RNase H assays were performed as described previously with 15 µg of RNA (Lemay *et al.* 2010). Briefly, total RNA was RNAse H-treated in a mixture containing a snoRNA-specific complementary oligonucleotide to induce cleavage and release a

specific 3’ ₋, as well as with or without oligo d(T) to remove the heterogeneity of poly(A) tails. Preparation of probes, blot hybridization, washes, and quantification of signals were done as described previously (Lemay *et al.* 2016).

2.2.10.3.5. Chromatin immunoprecipitation (ChIP) assays

ChIP-qPCR and ChIP-seq experiments were performed as described previously (Lemay *et al.* 2016). Chromatin was immunoprecipitated directly using Pan Mouse IgG Dynabeads (Life Technologies, 11041) for TAP-tagged proteins. ChIP were also performed using Pan Mouse IgG Dynabeads coated with antibodies specific to Rpb1 (clone 8WG16; Convance, MMS-126R) or the hemagglutinin (HA) sequence (clone 12CA5, Roche, 11 583 816 001). ChIP with phospho-specific CTD antibodies were performed using 2 µg of antibody: Ser2-P (Abcam, ab5095) and Ser5-P (Abcam, ab5131) were coated to Dynabeads M-280 Sheep anti-rabbit IgG beads (Life Technologies, 11203D) whereas Tyr1-P (clone 3D12, Active Motif, 61383) and Ser7-P (clone 4E12, EMD Millipore, 04-1570) were coated to Dynabeads Protein G beads (Life Technologies, 10003D). Control ChIP assays with untagged strains or with isotype matched control antibody were performed throughout the study.

ChIP-seq profiles of total RNAPII (Rpb1; clone 8WG16) in WT, *nab3Δ*, *sen1Δ*, *dbl8Δ* and *sen1Δ*/*dbl8Δ* strains were performed in YES media as described previously (Lemay *et al.* 2016). ChIP-seq experiments of total RNAPII and phospho-CTD modifications in wild-type and *dhp1*-depleted cells were performed with a spike-in adjustment procedure that allows a quantitative comparison between independent strains or conditions (Orlando *et al.* 2014). Chromatin preparations of *S. pombe* FBY1995 (Rpb3-HA; WT) and FBY2016 (Rpb3-HA; P_{nmt1}-*dhp1*) strains and *S. cerevisiae* FBY2064 strain (reference strain containing a copy of HA-tagged histone

H2B) were done as a regular ChIP, except that chromatin extracts have been frozen in liquid nitrogen and stored at -80°C before affinity purifications. For Tyr1-P, Ser2-P, Ser5-P, Ser7-P, and 8WG16 IPs, a *S. pombe*/*S. cerevisiae* chromatin ratio of 0.9:0.1 was used, whereas a chromatin ratio of 0.995:0.005 was used for Rpb3-HA purifications. Following combination of *S. pombe*/*S. cerevisiae* chromatin extracts, ChIP assays were performed as described above using antibody-coated beads. ChIP-seq profiles of total RNAPII from WT and *ysh1-AA* strains were also performed with a spike-in adjustment (ratio *S. pombe*/*S. cerevisiae* of 0.9:0.1) using extracts from *S. pombe* that were previously grown in EMM and treated for 4 hours with 2.5 µg/ml of rapamycin. Similarly, ChIP-qPCR of Rpb1 in *ysh1* nuclease mutants (Fig. 14) were grown in EMM and treated with rapamycin for 4h. Growth defects were not observed between WT and *ysh1-AA* strains after 4h-15h of rapamycin treatment as determined by kinetics growth curves using a Powerwave HT microplate spectrophotometer from Biotek (data not shown).

2.2.10.3.6. Protein analysis

Protein analysis was essentially performed as described (Lemay *et al.* 2016). Briefly, total extracts were prepared by lysis of mid-log phase *S. pombe* cultures in ice-cold lysis buffer (50 mM Tris-HCl (pH 7.5), 5 mM $MgCl_2$, 150 mM NaCl and 0.1 % NP-40 supplemented with 1 mM PMSF and 1x PLAAC) using a Precellys 24 homogenizer system (Bertin Technologies). Clarified lysates were normalized for total protein concentration using Bradford protein assay. Total extract proteins were separated by SDS-PAGE, transferred to nitrocellulose membranes, and analyzed by immunoblotting using antibodies against the hemagglutinin (HA) protein (Roche, 11 583 816 001; 1:500 (v/v) dilution), the GFP protein (11 814 460 001; 1:500 (v/v), the protein A tag (Sigma-Aldrich, P3775; 1:10000 (v/v) dilution), the FLAG peptide (Sigma-Aldrich,

F1804; 1:500 (v/v) dilution) and the α-tubulin protein (Sigma-Aldrich, T5168; 1:1000 (v/v) dilution). Membranes were then probed with goat anti-rabbit or anti-mouse secondary antibodies conjugated to IRDye 800CW (LI-COR, 926-32213; 1:15000 (v/v) dilution) and AlexaFluor 680 (Life Technologies, A-21057; 1:15000 (v/v) dilution), respectively. Protein detection was performed using an Odyssey infrared imaging system (LI-COR).

2.2.10.3.7. *snR99* constructs with different terminators

snR99 constructs containing *snR99*, *trx1* or *rps2* terminator sequences were generated using the *ade6* integration vector (pFB366) containing 1-kb of *snR99*-specific 5' promoter sequences followed by the *snR99* snoRNA region. The *snR99* snoRNA region was then fused to either 1-kb of *snR99* 3' sequences (pFB600), of *trx1* 3' sequences (pFB622), and *rps2* 3' sequences (pFB652). All constructs were confirmed by DNA sequencing. For single integration into the *ade6* locus, pFB600, pFB622 and pFB652 were linearized by digestion with the *AatII* restriction enzyme and transformed into a *snR99*-null strain. Positive integrants were confirmed by growth selection on EMM agar plates lacking adenine and expression were confirmed by RT-qPCR using primers located in the *snR99* region. Deletion of *pab2* was than generated in these strains using the kanMX6 marker.

2.2.10.3.8. Ysh1 expression constructs

The wild-type Ysh1 expression constructs was created by a 3-steps cloning procedure using the *ade6* integration plasmid (pFB366) as a host vector containing 640-bp of *ysh1*

5’ UTR sequences, the *ysh1* coding sequence including introns, and 846-bp *ysh1* 3’ UTR sequences. Fusion of a 3x FLAG tag to C-terminus of Ysh1 was performed using Q5 site-directed mutagenesis using primers containing the 3x FLAG tag DNA sequence, generating plasmid pFB1337. To generate amino acid substitutions at residies involved in Ysh1 endonuclease activity, histidine (H) residues 165 or 403 of Ysh1 were substituted to phenylalanine (F) to create pFB1355 (H165F) and pFB1358 (H403F), respectively. All mutations were performed by site-directed mutagenesis using pFB1337 as a template. All DNA constructs were confirmed by DNA sequencing. For single integration into the *ade6* locus, pFB366, pFB1337, pFB1355, and pFB1358 were linearized and transformed into FBY2066 and/or FBY2110 strains. Positive integrates were confirmed by growth selection on EMM agar plates lacking adenine and leucine and by western blotting.

2.2.10.3.9. Library preparation and Illumina sequencing

DNA librairies for ChIP-seq experiments were prepared as described previously (Lemay *et al.* 2016) using either the NEBNext ChIP-Seq Library Prep Master Mix Set for Illumina kit (New England BioLabs) or the SPARK DNA Sample Prep Kit Illumina Platform (Enzymatics) according to the manufacturer’s instructions.

2.2.10.3.10. ChIP-Seq processing

Briefly, the raw reads were trimmed using Trimmomatic version 0.32 (Bolger *et al.* 2014) with param ILLUMINACLIP:2:30:15 LEADING:30 TRAILING:30 MINLEN:23, and quality inspection was conducted using FastQC version 0.11.4 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Considering that most of

the samples were generated from chromatin contain exogenous spiked-in from *S. cerevisiae*, the trimmed reads from all datasets were aligned using BWA version 0.7.12-r1039 (Li and Durbin 2010) with the algorithm `mem` or `aln` depending on the read length (above or below 70 nt) and default parameters, onto a concatenated reference genome containing both the sequences of the *S. pombe* ASM294v2 and *S. cerevisiae* sacCer3 assemblies, inpired by others (Orlando *et al.* 2014). Note that no filtering on mapQ was performed in order to avoid discarding the signal at regions of the genome that are duplicated (BWA is randomly assigning the reads), but we generated mappability tracks for various read length to help the interpretation of particular regions. The number of reads mapped on the concatenated genomes varied from 91.5 % to 99.6 % of the sequenced reads (Table S1). Note that <0.28 % of reads coming from samples without spike-in were aligning to the *S. cerevisiae* genome (Table S1). Signal density files in BedGraph format were then generated using BEDTools genomecov version 2.17 (Quinlan and Hall 2010) with default parameters, then converted in uniform 10 nt bins WIG files for further normalisation steps (inspired by the script bedgraph_to_wig.py [https://gist.github.com/svigneau/8846527]).

2.2.10.3.11. ChIP-Seq normalization

Each signal density file of the datasets without spike-in was scaled such that the total sum of the signal over the *S. pombe* genome is equivalent to 1M reads of 100 nt, then the signal of the input dataset was subtracted from its corresponding IP dataset to generate the "sclWT-ctrl" files. The normalization of the datasets with spike-in was inpired by the method of presented by Orlando *et al.* (Orlando *et al.* 2014). For each dataset, the sum of the signal coming from the reads aligned to the *S. cerevisiae* genome was first scaled to the equivalent of 1M mapped reads of 100 nt, and the same scaling factor was applied to the signal coming from the reads aligned to the *S. pombe* genome to generate the "normSI" files. For the datasets generated in a WT and mutant

strains, the *S. pombe* signal was next scaled to a total signal equivalent to 1M mapped reads of 100 nt and the same scaling factor was applied to the corresponding mutant dataset to generate the “normSI_sclWT” files. The signal of the input dataset was then subtracted from its corresponding IP to generate the final “normSI_sclWT-ctrl” files used in downstream analyses.

The normalized WIG files were then encoded in bigWig format using the Kent utilities (Kent *et al.* 2010) Visual inspection of the data was performed using an AssemblyHub on the UCSC Genome Browser (Casper *et al.* 2018).

https://genome.ucsc.edu/cgi-bin/hgHubConnect?hubUrl=https://datahub-i8kms5wt.udes.genap.ca/CTD_sno_pombe/hub.txt&hgHub_do_firstDb=on&hgHub_do_redirect=on&hgHubConnect.remakeTrackHub=on

2.2.10.3.12. 3’READS analysis

The polyA sites (PAS) data from 3’READS experiments generated in (Liu *et al.* 2017) were reannotated to assign each PAS to the closest gene, including ncRNAs. The GTF genome annotations file version ASM294v2.29 was downloaded from PomBase (Wood *et al.* 2012; McDowall *et al.* 2015) in May 2017. However, based on manual inspection of our ChIP data, the orientation of three snoRNAs (namely SPSNORNA.36, SPSNORNA.37 and SPSNORNA.41) was inverted. In more details, PAS supported with at least two reads outside CDS were associated to the closest 3’ end gene annotation in a strand-specific manner within 1kb, whereas PAS inside CDS were ignored (except for overlapping genes if the closest 3' end was within 500 pb). For each gene, the strongest PAS identified in the WT strain grown in minimal medium was used to modify the transcriptional 3’ end gene coordinate provided for mRNA genes (Table S2) and snoRNA genes (Table S3).

2.2.10.3.13. ChIP-seq average profiles

The Versatile Aggregate Profiler (VAP) tool (Coulombe *et al.* 2014; Brunelle *et al.*, 2015) version 1.1.0 was used to generate the average profiles with the following common parameters: Annotation mode, Absolute analysis method, 10 bp windows size, mean aggregate value, smoothing of 6, and missing data were considered as "0" (with the exception of Fig. 9A where windows size was set to 50 pb). We used 4 references points for most of the figures to avoid contamination from adjacent genes; in such a case there are five blocks of data corresponding respectively to the upstream gene (signal ignore), the upstream intergenic region (signal aligned toward the TSS of the gene of interest), the gene of interest (signal split such that the first half was aligned toward the TSS and the second half toward the 3'end/polyA), the downstream intergenic region (signal aligned toward the 3'end/polyA), and the downstream gene (signal ignore). For figures showing either the average profile of datasets over the complete 4,755 mRNA genes with associated PAS (Fig. 9A and 12A), the 31 monocistronic snoRNA genes (Fig. 9B) or the 24 monocistronic snoRNA genes with identified PAS (Fig. 12B), only signal in the first and last blocks were ignored. Similarly, for figures focusing only on the 3' region of the 4,755 mRNA genes with associated PAS (Fig. 9F and 12E), the 31 monocistronic snoRNA genes (Fig. 9G) or the 24 monocistronic snoRNA genes with identified PAS (Fig. 12F), only the signal in the second half of the third block (corresponding to the gene of interest) and the fourth block (downstream intergenic region) was shown. To better represent the extension of signal in the mutant strains (Fig. 13D-G, 14D-E, S4A and S4K-L), we used only 2 reference points where the first block contain the signal in the upstream intergenic region potentially contaminated with signal from the upstream gene (signal aligned as described previously), the second block contains the signal of the gene of interest, and the third block the downstream intergenic region also potentially contaminated with signal from the downstream gene.

Genome-wide Pearson correlation coefficients (Fig. 12G and S4) were calculated using the epiGeEC tool version 1.0 (Laperle *et al., submitted*).

2.2.10.4. Supplementary references

Bahler J, Wu JQ, Longtine MS, Shah NG, McKenzie A, 3rd, Steever AB, Wach A, Philippsen P, Pringle JR. 1998. Heterologous modules for efficient and versatile PCR-based gene targeting in *Schizosaccharomyces pombe*. *Yeast* **14**: 943-951.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Brunelle M, Coulombe C, Poitras C, Robert MA, Markovits AN, Robert F, Jacques PE. 2015. Aggregate and Heatmap Representations of Genome-Wide Localization Data Using VAP, a Versatile Aggregate Profiler. *Methods Mol Biol* **1334**: 273-298.

Casper J, Zweig AS, Villarreal C, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, Lee CM, Lee BT, Karolchik D et al. 2018. The UCSC Genome Browser database: 2018 update. *Nucleic Acids Res* **46**: D762-D769.

Coulombe C, Poitras C, Nordell-Markovits A, Brunelle M, Lavoie MA, Robert F, Jacques PE. 2014. VAP: a versatile aggregate profiler for efficient genome-wide data representation and discovery. *Nucleic Acids Res* **42**: W485-493.

Ding L, Laor D, Weisman R, Forsburg SL. 2014. Rapid regulation of nuclear proteins by rapamycin-induced translocation in fission yeast. *Yeast* **31**: 253-264.

Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**: 2204-2207.

Lemay JF, D'Amours A, Lemieux C, Lackner DH, St-Sauver VG, Bahler J, Bachand F. 2010. The nuclear poly(A)-binding protein interacts with the exosome to promote synthesis of noncoding small nucleolar RNAs. *Mol Cell* **37**: 34-45.

Lemay JF, Larochelle M, Marguerat S, Atkinson S, Bahler J, Bachand F. 2014. The RNA exosome promotes transcription termination of backtracked RNA polymerase II. *Nat Struct Mol Biol* **21**: 919-926.

Lemay JF, Marguerat S, Larochelle M, Liu X, van Nues R, Hunyadkurti J, Hoque M, Tian B, Granneman S, Bahler J *et al.* 2016. The Nrd1-like protein Seb1 coordinates cotranscriptional 3' end processing and polyadenylation site selection. *Genes Dev* **30**: 1558-1572.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Liu X, Hoque M, Larochelle M, Lemay JF, Yurko N, Manley JL, Bachand F, Tian B. 2017. Comparative analysis of alternative polyadenylation in *S. cerevisiae* and *S. pombe*. *Genome Res* **27**: 1685-1695.

McDowall MD, Harris MA, Lock A, Rutherford K, Staines DM, Bahler J, Kersey PJ, Oliver SG, Wood V. 2015. PomBase 2015: updates to the fission yeast database. *Nucleic Acids Res* **43**: D656-661.

Orlando DA, Chen MW, Brown VE, Solanki S, Choi YJ, Olson ER, Fritz CC, Bradner JE, Guenther MG. 2014. Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell reports* **9**: 1163-1170.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

Waddle JA, Karpova TS, Waterston RH, Cooper JA. 1996. Movement of cortical actin patches in yeast. *J Cell Biol* **132**: 861-870.

Wood V, Harris MA, McDowall MD, Rutherford K, Vaughan BW, Staines DM, Aslett M, Lock A, Bahler J, Kersey PJ *et al.* 2012. PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res* **40**: D695-699.

CHAPITRE 3
MOCKSCREEN : ÉVALUATION DE LA FRÉQUENCE DU NOMBRE DE COUPLES À RISQUE D'AVOIR UN ENFANT ATTEINT D'UNE MALADIE MENDÉLIENNE RÉCESSIVE RARE

Ce chapitre cherche à évaluer le nombre de couples aléatoires pouvant engendrer un enfant à risque de développer une maladie génétique récessive rare. En collaboration avec le groupe de recherche de Sébastien Lévesque, médecin-généticien au Centre hospitalier universitaire de Sherbrooke, des milliers de fichiers contenant l'annotation de centaines de milliers de variants génétiques seront analysés pour comparer deux méthodologies : utiliser les fichiers de variants de vraies personnes et utiliser les fréquences populationnelles des variants pour déterminer la fréquence d'apparition de couples à risque. La comparaison nous permet d'évaluer si ces deux approches sont équivalentes.

3.1. Matériel et méthode

Cette section décortique les éléments importants de l'usage et l'implémentation de MockScreen. L'implémentation de MockScreen est faite en python3.4.0 et prend en entrée un mode d'analyse (0 : générer de nouveaux couples, 1 : utiliser une liste de couples fournie), la liste de VCF annotée par snpEFF (version 4.3), la liste contenant le sexe de chaque individu, le nombre de couples et de groupes désirés, ainsi qu'un préfixe pour l'analyse. MockScreen utilise une copie interne et adaptée des bases de données ExAC (version 0.3.1) et ClinVar (version 20180225).

3.1.1. Données utilisées

Pour ce projet, plusieurs milliers d'individus sont nécessaires afin de pouvoir offrir des résultats crédibles et solides. Pour ce faire, le *National Institutes of Health* (NIH) est un bon candidat pour nous fournir le matériel brut à une analyse de cette ampleur. En effet, le NIH consiste à l'institut Américain le plus important dans le domaine de la santé, où plusieurs études ont pu générer des échantillons de WES de milliers de personnes. Par contre, ce genre de données ne s'accède pas facilement : obtenir l'approbation du comité d'éthique institutionnel quant à la validité du projet, ouvrir un compte NIH, remplir une demande expliquant la nature du projet (et ce pour chaque cohorte avec des paramètres différents), fournir les documents d'acceptation éthiques, obtenir les approbations certifiant la sécurité des données, télécharger les données sous accès contrôlé, et finalement les décrypter pour commencer les analyses. Les résultats qui suivent ne sont qu'une preuve de concept étant donné que pour l'instant, je n'ai accès qu'à 1968 individus indépendants (653 femmes et 1315 hommes) faisant partie de la cohorte *Ottawa Heart Study* (phs000806.v1.p1), mise en place pour étudier les problèmes cardiaques. Les données de variants génétiques de la cohorte sont toutes regroupées dans un seul gros fichier de format VCF contenant plus de 6M de lignées et 1977 colonnes.

3.1.2. Filtres utilisés

La première étape de MockScreen est de sélectionner les variants rares ayant un fort potentiel d'impact fonctionnel dans les gènes connus pour être impliqués dans des maladies récessives puisqu'ils sont les seuls intéressants pour notre étude. Une fois sélectionné, un variant est attribué à l'intérieur de MockScreen aux individus porteurs de ce variant. Une autre série de filtres s'appliquera ainsi aux informations de l'individu.

3.1.2.1. Sélection des variants

Pour sélectionner les variants rares, la valeur de fréquence allélique de ExAC doit être inférieure à 1 %. Dans le cas où le variant n'a aucune valeur pour ExAC, sa valeur de CAF doit être inférieure à 2 %. Lorsqu'aucune de ces valeurs n'est disponible, nous considérons automatiquement qu'il est rare.

Pour sélectionner les variants pathogéniques (*Pathogenic*) ou potentiellement pathogéniques (*Likely pathogenic)*, la base de données ClinVar est utilisée. Les variants présents dans ClinVar mais classés autrement (e.g. *Benign*) sont automatiquement rejetés. Pour les variants absents de ClinVar, l'annotation de snpEff est utilisée pour estimer leur impact ; seulement les impacts notés élevés (*High*) et modérés (*Medium*) ont été gardés. À noter que les variants gardés selon ClinVar ont tous été mis dans le groupe à impact élevé, peu importe la prédiction de snpEff.

La dernière étape concerne la nécessité pour un variant d'être inclus dans un des 450 gènes connus pour être impliqués une maladie génétique rare. Cette sélection a été préparée par l'équipe de Sébastien Lévesque et est présentée à l'Annexe 1.

3.1.2.2. Sélection des porteurs

Puisque nous nous intéressons aux maladies récessives, seuls les individus hétérozygotes pour un variant sont conservés pour notre analyse. Pour chaque individu

et chaque variant, le fichier VCF contient le statut d'hétérozygocité prédit ; la notation qui nous intéresse est 0/1 – 0 indiquant l'allèle référence et 1 l'allèle alternative. De plus, pour qu'un individu soit considéré comme un porteur du variant, il doit avoir un minimum de couverture de 10 (ou en d'autres mots, au moins 10 lectures supportant ce génotype). Le ratio des lectures contenant l'allèle alternative doit aussi être entre 15 % et 85 %, sinon le génotype de cet individu est considéré comme homozygote.

3.1.3. Recensement des variants présents

MockScreen génère des fichiers contenant les statistiques sur les variants de la population et les gènes affectés de la cohorte à l'étude. Le tableau résumant les variants contient pour chaque variant son impact, son gène touché, le nombre d'individus porteur et sa fréquence dans la cohorte étudiée (e.g. Annexe 2). Le tableau résumant les gènes d'intérêt contient pour chaque gène, le nombre d'individus porteurs d'au moins un variant inclus dans ce gène pour chaque type d'impact, la somme de tous les variants inclus dans ce gène, le nombre différent de variants inclus dans ce gène et la fréquence de la population ayant au moins un variant inclus dans ce gène (e.g. Annexe 3).

3.1.4. Définition d'un couple à risque

Étant donné que deux types d'impacts de variants sont acceptés lors des filtres (élevé et modéré), il est possible d'être plus ou moins sévère dans la définition d'un couple à risque. En effet, le niveau de risque un couple dont les deux individus possèdent un

variant d'impact élevé n'est pas le même que si les deux individus d'un couple possèdent un variant d'impact modéré. Pour cette raison, cinq groupes de couples à risque ont été définis :

- HH sont les couples dont chaque individu a au moins un variant d'impact élevé dans le même gène.
- HM sont les couples avec un individu ayant au moins un variant d'impact élevé et l'autre avec au moins un variant d'impact moyen dans le même gène.
- MM sont les couples dont chaque individu a au moins un variant d'impact moyen dans le même gène.
- HHM est une union des groupes HH et HM.
- HHMM est une union des groupes HH, MM et HM.

Il est à noter que l'attribution d'un couple à un groupe de risque n'est pas exclusive. Par exemple, un couple dont le père a un variant d'impact élevé dans un gène et que sa partenaire a deux variants dont un d'impact élevé et l'autre moyen dans le même gène, ferait partie des groupes de couples à risque HH et HM. Donc, il est attendu que la somme des fréquences des groupes HH et HM n'égal pas celle de HHM et que la somme des fréquences des groupes HH, HM et MM n'égal pas celle de HHMM.

3.1.5. Analyse des couples

Tous les individus avec leur liste de variants sont utilisés pour générer des couples. La création des couples est aléatoire. Plusieurs séries de couples sont générées où tous les individus sont remis entre chaque série ; il est donc possible qu'un même couple se forme dans deux séries distinctes (couple dupliqué). Le nombre de couples et le

nombre de séries sont paramétrés, alors il est facile de relancer l'analyse avec d'autres valeurs.

Une fois tous les couples créés, la fréquence de couples à risque de chaque série (calculée en divisant le nombre de couples à risque sur le nombre de couples créés) est utilisée pour faire une distribution représentée sous forme de diagramme de quartiles (*box plot*).

Cette analyse a été faite indépendamment pour tous les groupes de couples à risque. La reproductibilité des résultats avec 500 groupes de 500 couples a été vérifiée en effectuant deux itérations utilisant ces paramètres. De plus, une troisième itération a été effectuée en utilisant les mêmes individus, mais en leur attribuant des sexes aléatoires, afin de déterminer l'impact de l'attribution des sexes sur l'analyse.

3.1.6. Fréquence de couples à risque attendue

La fréquence attendue pour chaque groupe de couples à risque est calculée à partir des fréquences alléliques provenant des individus de la cohorte étudiée. Cette restriction permet d'éliminer le biais qui aurait été causé par l'utilisation d'une population différente lors de la comparaison des fréquences attendues et obtenues par les couples. Voici la démarche détaillée du calcul de la fréquence des couples à risque HH à titre d'exemple :

Pour chacun des gènes d'intérêt, la probabilité (p_g) d'avoir un couple à risque du groupe HH est calculée à partir du nombre d'individus (n_i) possédant au moins un variant d'impact élevé désiré, sur le nombre d'individus dans la cohorte étudiée (N). Ainsi, p_g est calculé en multipliant la probabilité d'avoir un individu possédant au moins un variant d'impact élevé désiré (p_{I1}) par la probabilité d'avoir un deuxième individu (p_{I2}). À noter que tout dépendant du groupe de couples à risque, cette démarche peut légèrement varier pour répondre aux conditions définies.

$$p_g = p_{I1} * p_{I2} = \frac{n_i}{N} * \frac{n_i - 1}{N - 1}$$

Comme la fréquence attendue ($F_{attendue}$) de couples à risque est égale à la probabilité d'avoir au moins un gène à risque, il est plus facile de soustraire à 1 la probabilité de ne pas avoir de gène à risque ($F_{inverse}$). Pour obtenir ($F_{inverse}$), il suffit de multiplier la probabilité de chaque gène de ne pas être à risque :

$$F_{attendue} = 1 - F_{inverse} = 1 - \prod_{g=0}^{x}(1 - p_g)$$

où x correspond au nombre de gènes étudiés.

3.1.7. Comparaison des distributions des fréquences obtenues aux fréquences attendues

Afin de comparer la fréquence attendue avec la distribution des fréquences obtenues par les séries de couples de la première itération, la moyenne et l'écart-type de cette

distribution sont utilisées pour déterminer si elle suit une loi normale, pour ensuite, déterminer si la moyenne est significativement différente de la fréquence attendue en appliquant le test statistique de Student. Ce test n'est valable que pour une distribution normale. Cette approche est appliquée à chaque groupe de couples à risque.

3.2. Résultats

Bien que le nombre d'individus disponibles ne soit probablement pas assez élevé pour évaluer de façon confiante la fréquence de couples à risque dans la population générale, il est suffisant pour évaluer l'impact d'utiliser des couples comparés à utiliser directement la fréquence attendue pour la cohorte étudiée.

Il est à noter que quelques détails techniques de MockScreen restent à régler : les variants possédant plusieurs allèles alternatifs sont tronqués pour garder que le premier et un variant impliqué dans plus qu'un gène d'intérêt n'est considéré comme impliqué que dans le premier de ceux-ci.

3.2.1. Recensement des variants présents

Les 1968 individus ont collectivement plus de 6M de variants, mais seulement 5306 variants rares potentiellement pathogéniques (806 d'impact élevé et 4500 d'impact modéré) ont réussi à passer les filtres et ce, dans 410 des 450 gènes du panel.

Ce module de MockScreen, a permis de raffiner et définir nos filtres utilisés. En effet, lors de nos premiers essais du code, plusieurs variants se trouvaient dans plus de 1 % de la population. En regardant en détail ces variants (grâce aux tabulateurs générés par ce module), il était clair que la majorité des variants n'étaient pas rares ni pathogéniques. Les filtres, étant trop permissifs, ont été modifiés en identifiant des points communs des variants clairement aberrants. L'utilisation de cette approche a été nécessaire plus d'une fois afin d'arriver aux filtres finaux définis dans la section 3.1.2. Maintenant, seuls 31 variants (2 d'impact élevé et 29 d'impact modéré) se trouvent dans plus de 1 % de la population (annexe 2). À titre d'exemple, avant l'ajout de la dernière modification des filtres (enlever les variants notés non pathogéniques dans ClinVar), 164 variants (15 d'impact élevé et 149 d'impact modéré) se trouvaient dans plus de 1 % de la population.

3.2.2. Analyse des couples

Pour la première itération, 500 groupes de 500 couples ont été générés. Bien que 14,04 % (71,25 par série) des couples générés fassent partie du groupe de couples à risque HHMM (voir section 3.1.3), seulement 0,34 % (1,7 couple par série) des couples générés font partie du groupe HH (Tableau 3). Le groupe MM est de loin le plus fréquent, ce qui était attendu étant donné que les variants d'impact moyen étaient fortement plus représentés que ceux d'impact élevé. Si on enlève les couples du groupe MM du groupe HHMM (soit le groupe HHM), nous obtenons une fréquence de couples à risque de 2,23 % (11,15 couples par série).

La comparaison entre les deux premières itérations suggère une bonne reproductibilité des résultats avec 500 groupes de 500. En effet, le test statistique de Wilcoxon montre que les distributions des fréquences de couples à risque ne sont pas différentes entre

les deux itérations et ce pour chaque groupe de couples à risque (Figure 16). De plus, comme pour les deux premières itérations, aucune différence significative entre les distributions avec les bons sexes et la distribution avec les sexes aléatoires n'a été détectée (Figure 16). Ces résultats suggèrent que les paramètres utilisés sont suffisants pour fournir des résultats reproductibles pour cette population d'individus et que les sexes des individus ne semblent pas avoir d'impact réel sur l'évaluation de la fréquence de couples à risque. Il est à noter que seuls les variants sur les chromosomes autosomaux ont été étudiés.

3.2.3. Fréquences de couples à risque attendues

Le Tableau 4 contient les fréquences attendues (calculées à partir des fréquences alléliques) pour chaque groupe de couples à risque. Tel qu'attendu, les groupes de couples à risque nécessitant au moins un variant d'impact élevé (HH, HM et HHM) ont une plus faible fréquence attendue que les autres groupes de couples à risque (MM et HHMM) étant donné que dans la population, seulement une faible proportion des variants a un impact élevé.

3.2.4. Comparaison des distributions des fréquences obtenues aux fréquences attendues

Pour vérifier l'impact de faire des couples à partir de fichiers VCF, les distributions de fréquences obtenues des différents groupes de couples à risque (Tableau 3) ont été comparées avec leur fréquence attendue correspondante (Tableau 4). À première vue, l'ordre de grandeur et la tendance sont similaires entre les deux approches, ce qui fournit un bon contrôle positif sur la justesse de l'algorithme d'analyse utilisé. Pour

comparer statistiquement les moyennes à l'aide du test de Student, il est nécessaire que les distributions sous-jacentes suivent une loi normale, ce qui n'est pas le cas avec la distribution des fréquences obtenues de couples à risque du groupe HH (Tableau 5). Une faible diversité des couples répondant aux critères d'un couple à risque HH, causé par le manque d'individu, explique ce manque de normalité de la distribution. Bien que l'augmentation des individus n'augmente pas le nombre de variants d'impact élevé de façon relative, l'augmentation absolue de ce type de variants permettrait plus de combinaisons différentes de couples à risque d'impact HH.

Tableau 3. Moyenne et écart-type des distributions des fréquences des différents groupes de couples à risque obtenus pour les 500 séries et trois itérations.

Groupe de couples à risque	Itération	Moyenne	Écart-type
HH	Itération 1	0,34%	0,26%
	Itération 2	0,35%	0,26%
	Sexes aléatoires	0,36%	0,26%
HM	Itération 1	1,91%	0,62%
	Itération 2	1,94%	0,61%
	Sexes aléatoires	1,90%	0,63%
MM	Itération 1	12,16%	1,40%
	Itération 2	12,17%	1,37%
	Sexes aléatoires	12,28%	1,41%
HHM	Itération 1	2,23%	0,67%
	Itération 2	2,27%	0,67%
	Sexes aléatoires	2,23%	0,67%
HHMM	Itération 1	14,04%	1,50%
	Itération 2	14,07%	1,49%
	Sexes aléatoires	14,14%	1,51%

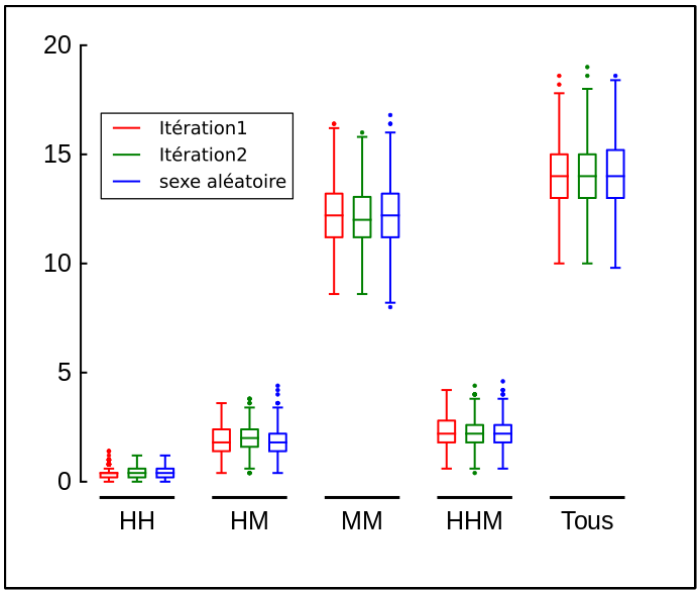


Figure 16. Distribution des fréquences des différents groupes de couples à risque pour trois itérations différentes.
Aucune différence significative entre les trois itérations pour chaque groupe de couples à risque (test Wilcoxon *signed-rank*).

Tableau 4. Fréquences attendues pour chaque groupe de couples à risques

Couples HH	Couples HM	Couples MM	Couples HHM	Couples HHMM
0,36%	0,96%	12,25%	1,31%	14.13%

Concernant les groupes ayant une distribution de fréquence obtenue normale, les moyennes obtenues et attendues des groupes MM et HHMM sont similaires, alors que pour les groupes HM et HHM l'utilisation des fréquences attendues (calculées par les fréquences alléliques) donne des résultats significativement plus bas que les fréquences obtenues par les profils de mutations individuels (VCF). Ce résultat nous semble cohérent avec le fait que les groupes de couples à risque contenants deux variants d'impact modéré sont potentiellement ceux possédant le plus de faux positifs,

alors que l'utilisation des VCF pour les groupes de couples à risque HM et HHM apporte une meilleure précision.

Tableau 5. Test de normalité et de Student sur la comparaison de moyennes de distributions des fréquences de couples à risque de chaque groupe de l'itération1

	HH	HM	MM	HHM	HHMM
Valeur ρ, Test de normalité*	1,54E-15	0,017	0,390	0,011	0,275
Valeur ρ, Test de Student de comparaison de moyennes**	N/A	1,07E-133	0,142	4,60E-117	0,161

*H_0 : La distribution suit une loi normale, rejetée si ρ<0,01

**H_0 : La moyenne de la distribution est semblable à la moyenne attendue, rejetée si ρ<0.01

3.3. Discussion

L'implémentation modulable et paramétrable de MockScreen a permis de mettre sur pied une preuve de concept avec un nombre restreint d'individus. Cette preuve de concept a permis un avancement important sur la mise au point de plusieurs filtres rendant MockScreen prêt pour être utilisé sur une plus grande population.

De plus, la séparation des différents groupes de couples à risque permet d'avoir un éventail de résultats permettant de mieux conceptualiser et analyser la fréquence de couples à risque d'une population étudiée en fonction la sensibilité désirée. En effet, probablement que les gènes du groupe MM, soient avec deux allèles d'impacts modérés, n'est pas assez problématique pour être pris en compte pour ce genre d'analyse.

L’ajustement paramétré du nombre de groupes et le nombre de couples à générer permettent rapidement de tester les paramètres optimaux pour avoir une bonne reproductibilité dépendamment du nombre d’individus analysés. Ces tests ont aussi permis de montrer que l’attribution des sexes des individus n’est pas une étape critique pour cette analyse. Ce détail permettra de considérer des cohortes pour lesquelles les sexes des individus ne sont pas disponibles.

Cette preuve de concept a permis de confirmer que l’utilisation des couples était une bonne approche et que le calcul des fréquences attendues par l’utilisation des fréquences alléliques semble une approche raccourcie biologiquement erronée.

CHAPITRE 4.
DISCUSSION ET CONCLUSION GÉNÉRALE

L'objectif de ce mémoire était de développer et d'utiliser des outils bio-informatiques dans le cadre de deux collaborations, soit un projet en génomique et un projet en génétique.

Dans le cadre du projet en génomique, l'outil de normalisation SpkNorm a été développé dans le but d'analyser des données de ChIP-Seq-SI. L'analyse d'une quarantaine d'ensembles de données a permis de tester différentes méthodes et paramètres pour l'optimisation de SpkNorm. Basés sur ces résultats et sur d'autres expériences effectuées, nous avons montré que contrairement à ce qui est retrouvé chez *S. cerevisiae*, la terminaison de la transcription des gènes non-codants par l'ARN polymérase II chez *S. pombe* est causée par le modèle *Torpedo.* La mise au point d'une version de SpkNorm capable d'analyser d'autres organismes rendrait son utilisation plus polyvalente. En attendant, SpkNorm reste un bon outil pour analyser des données de ChIP-Seq-SI de *S. pombe*.

Dans le cadre du projet en génétique, l'outil modulaire et paramétrable MockScreen a été développé dans le but d'identifier la fréquence de couples à risque d'engendrer un enfant ayant une maladie récessive rare. Les résultats obtenus à partir d'un relativement faible nombre d'individus ont permis le développement de MockScreen et suggèrent que la génération de couples synthétiques est l'approche à prioriser contrairement à l'utilisation des fréquences alléliques populationnelles. MockScreen est maintenant prêt à analyser plusieurs milliers de fichiers génétiques provenant de différentes cohortes, afin de fournir un résultat solide et fournir ainsi une donnée cruciale pour la mise en relation des coûts de WES avec les impacts (financiers et autres) de ces maladies pour la société. Cette mise en relation touche cependant des aspects éthiques qui vont bien au-delà du présent mémoire.

ANNEXE 1.

GROUPES DE GÈNES ÉTUDIÉS POUR LE PROJET GÉNÉTIQUE

AAAS	CFI	DNAL4	GUCY1A3	LRRC6	OAT
AARS2	CFL2	DNASE1L3	GYG1	MALT1	OPA3
ABHD12	CHAT	DPAGT1	GYS1	MAN2B1	PAH
ACADS	CHRNA1	DPM2	HADHA	MCM4	PANK2
ACADVL	CHRNE	DSG1	HADHB	MEGF10	PCYT1A
ACP5	CLCN1	DST	HARS	MERTK	PDX1
ACTA1	CLCN2	DTNBP1	HARS2	MFSD8	PEX10
ADAR	CLDN16	DYNC2H1	HEPACAM	MLYCD	PEX2
ADCK3	CLPB	ECM1	HEXA	MOCS1	PFKM
AFG3L2	CNGA3	EGR2	HEXB	MOCS2	PGAM2
AGA	COASY	EIF2B1	HF1	MPL	PGM3
AGK	COL11A2	EIF2B2	HPCA	MPZ	PHKG2
AGRN	COL17A1	EIF2B3	HSD11B2	MRE11A	PLA2G6
AGXT	COL18A1	EIF2B4	HSD17B4	MS4A1	PLCE1
AIRE	COL2A1	EIF2B5	HYDIN	MTFMT	PLEC1
ALMS1	COL4A3	ELAC2	IER3IP1	MTO1	PLEKHG5
ALPL	COL4A4	ENPP1	IHH	MTPAP	PLG
ALS2	COL6A2	EPM2A	IL12RB1	MUT	PMP22
AMN	COL9A2	ERCC6	IL21R	MYH7	PNKP
ANKS6	COLQ	ERCC6L2	INPP5E	MYO1E	POC1A
ANTXR1	COQ2	ERLIN2	INPPL1	MYO5A	POLG
ANTXR2	COQ6	FBXO7	IQCB1	NADK2	POLR3A
APTX	CORO1A	FCGR3A	ISCA2	NAGA	POMGNT1
ARMC4	COX10	FCYT	ISCU	/NAGLU	POMT1
ARSA	COX15	FIG4	ISPD	NAGS	PPT1
ARSB	COX6A1	FKBP10	ITGB4	NALCN	PRICKLE1
B3GALT6	CPA6	FKRP	ITK	NBEAL2	PRKACG
B3GALTL	CPT2	FLVCR1	JAGN1	NCF4	PRKCD
B3GAT3	CR2	FOXRED1	KIAA0226	NDST1	PROS1
BCS1L	CTC1	FTL	KIF1A	NDUFA10	PRX
BSCL2	CTDP1	FUCA1	KRT14	NDUFA12	PSMB8
C10orf2	CTNS	GAA	KRT5	NDUFA2	PTPN14
C12orf65	CTSD	GABRG2	L2HGDH	NDUFA9	PTRF
C15orf41	CUBN	GALC	LAMA3	NDUFAF2	PUS1
C3	CYC1	GAN	LAMB3	NDUFAF6	PYGL
CA5A	D2HGDH	GBA	LAMC2	NDUFS3	RAB27A
CABP4	DAG1	GBA2	LARGE	NDUFS4	RAG1
CAPN3	DCAF17	GCDH	LDHA	NDUFS7	RAG2
CC2D1A	DDHD1	GDAP1	LEP	NDUFS8	RARS2
CCNO	DHCR7	GIF	LIPN	NEFL	RBCK1
CCT5	DHFR	GNAT1	LMNA	NEU1	RBP4
CD19	DLD	GNPTAB	LPIN1	NHLRC1	RDH12
CD81	DLL3	GOSR2	LRBA	NPC1	REEP2
CEP83	DNAJC3	GTPBP3	LRP5	NPR2	RIPK4

RLBP1
RNF168
RSPH1
RTEL1
SCN5A
SCN9A
SDHA
SERAC1
SERPINA1
SERPINB7
SETX
SGCA
SGCB
SLC12A3
SLC17A5
SLC19A2
SLC22A5
SLC25A19
SLC25A4
SLC2A1
SLC33A1
SLC34A1
SLC34A3
SLCO1B1
SLCO1B3
SMN1
SMN2
SMPD1
SOX18
SPG11
SPG20
SRD5A3
ST3GAL3
STT3B
STX11
SUCLA2
SURF1
TACSTD2
TBC1D24
TH
TJP2
TK2
TMEM15
TMEM165
TMEM67
TPI1
TPK1
TPM3
TPP1
TRAPPC11
TRDN
TRIM2
TRIM32
TSC1
TSEN2
TSEN54
TTC7A
TULP1
TYR
UBR1
UCHL1
VPS13B
VRK1
WDR73
WNK1
XPNPEP3
YARS2
ZBTB24
ZMYND10
ABCB7
ABCD1
AIFM1
ALAS2
ALG13
AMER1
AP1S2
AR
ARHGEF9
ARX
ATP2B3
ATP6AP2
ATP7A
ATRX
BRCA2
BRWD3
BTK
CDKL5
CHM
CLCN5
CLIC2
COL4A5
COL4A6
CYBB
DAX1
DKC1
DLG3
DMD
DXS423E
EBP
EMD
F8
F9
FHL1
FLNA
FMR1
FOXP3
FOXRED1
FRMD7
FTSJ1
GJB1
GK
GLA
GPC3
GPR143
HCCS
HPRT1
HSD17B10
IGSF1
IKBKG
KDM6A
KIF4A
L1CAM
MAOA
MBTPS2
MECP2
MED12
MTM1
NDP
NDUFA1
NDUFA11
NDUFAF1
NDUFAF2
NDUFAF3
NDUFAF4
NDUFAF5
NDUFB3
NDUFB9
NDUFS1
NDUFS2
NDUFS3
NDUFS4
NDUFS6
NDUFV1
NDUFV2
NHS
NLGN3
NLGN4
NSDHL
NUBPL
OCRL
OFD1
OTC
PCDH19
PDHA1
PHEX
PHF6
PHKA1
PIGA
PLP1
PLS3
PRPS1
RAB39B
RP2
RPGR
RS1
SAT1
SHOX
SHOXY
SLC16A2
SLC35A2
SLC6A8
SLC9A6
SMS
SOX3
SRPX2
SYN1
TAZ
TIMM8A
TNFSF5
TRAPPC2
UBA1
WAS
WDR45
WT1
ZDHHC15
ZNF81

ANNEXE 2.
VARIANTS DANS PLUS DE 1 % DE LA POPULATION PASSANT LES FILTRES DE MOCKSCREEN

Chr_Pos_Ref_Alt	Nom du gène associé	Nombre individus	Fréquence dans la population	Impact
11_71146886_C_G	DHCR7	37	1,88 %	H
22_42457056_C_T	NAGA	20	1,02 %	H
19_40900179_TTCCTCC_TTCC	PRX	120	6,10 %	M
15_89876827_TTGCTGCTGCTGC_TTGCTGCTGC	POLG	116	5,89 %	M
18_77475187_TGGA_TGGAGGA	CTDP1	67	3,40 %	M
6_56434712_G_T	DST	59	3,00 %	M
21_46875982_G_A	COL18A1	50	2,54 %	M
6_161128812_G_A	PLG	48	2,44 %	M
21_46876004_C_T	COL18A1	48	2,44 %	M
6_161152905_C_T	PLG	47	2,39 %	M
1_150482145_G_A	ECM1	44	2,24 %	M
10_50824106_C_T	CHAT	43	2,18 %	M
18_28934375_T_C	DSG1	35	1,78 %	M
2_47287925_C_A	TTC7A	33	1,68 %	M
4_178360811_G_T	AGA	33	1,68 %	M
15_89870178_CCCT_ACCT	POLG	30	1,52 %	M
1_183204831_C_G	LAMC2	29	1,47 %	M
1_214558119_T_G	PTPN14	27	1,37 %	M
18_28934374_A_G	DSG1	27	1,37 %	M
12_102190521_C_T	GNPTAB	26	1,32 %	M
18_28914021_G_T	DSG1	26	1,32 %	M

16_71220668_C_T	HYDIN	25	1,27 %	M
X_153363099_CGCGGCGGCG_CGCGGCG	MECP2	24	1,22 %	M
16_87925430_T_C	CA5A	23	1,17 %	M
9_98638328_C_A	ERCC6L2	22	1,12 %	M
1_207641950_C_T	CR2	21	1,07 %	M
6_56457044_T_C	DST	20	1,02 %	M
8_145151237_C_CAGGTGG	CYC1	20	1,02 %	M
10_90537910_G_T	LIPN	20	1,02 %	M
14_51382637_G_A	PYGL	20	1,02 %	M
21_46924416_CGGCCCCCCCGGCCCCCCA_C	COL18A1	20	1,02 %	M

ANNEXE 3.

EXEMPLE DE TABLEAU RÉSUMANT LES VARIANTS D'IMPACT ÉLEVÉ POUR QUINZE GÈNES D'INTÉRÊT

Nom du gène associé	Nombre individus	Fréquence dans la population	Nombre de SNP	Nombre de SNP Différent
DHCR7	44	2,24%	44	7
TYR	34	1,73%	34	14
PAH	30	1,52%	30	16
NAGA	24	1,22%	24	3
PLG	19	0,97%	19	1
ACADVL	16	0,81%	16	6
CLCN1	13	0,66%	14	4
CUBN	12	0,61%	12	9
POLG	12	0,61%	12	6
TRDN	12	0,61%	12	3
DYNC2H1	11	0,56%	11	10
CPT2	11	0,56%	11	3
SLC12A3	10	0,51%	10	8
GAA	10	0,51%	10	6
GYG1	10	0,51%	10	2

BIBLIOGRAPHIE

Baejen, C., Andreani, J., Torkler, P., Battaglia, S., Schwalb, B., Lidschreiber, M., Maier, K.C., Boltendahl, A., Rus, P., Esslinger, S., *et al.* (2017). Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes. Mol. Cell *66*, 38–49.e6.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput. Biol. *9*, e1003326.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837.

Belandia, B., Orford, R.L., Hurst, H.C., Parker, M.G., Sweep, F.C., Span, P.N., and Stunnenberg, H.G. (2002). Targeting of SWI/SNF chromatin remodelling complexes to estrogen-responsive genes. EMBO J. *21*, 4094–4103.

Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I.M., Herr, W., Hernandez, N., Delorenzi, M., *et al.* (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. Genome Res. *24*, 1157–1168.

Brunelle, M., Coulombe, C., Poitras, C., Robert, M.-A., Markovits, A.N., Robert, F., and Jacques, P.-É. (2015). Aggregate and Heatmap Representations of Genome-Wide Localization Data Using VAP, a Versatile Aggregate Profiler. In Methods in Molecular Biology (Clifton, N.J.), pp. 273–298.

Carroll, K.L., Pradhan, D.A., Granek, J.A., Clarke, N.D., and Corden, J.L. (2004). Identification of cis Elements Directing Termination of Yeast Nonpolyadenylated snoRNA Transcripts. Mol. Cell. Biol. *24*, 6241–6252.

Casañal, A., Kumar, A., Hill, C.H., Easter, A.D., Emsley, P., Degliesposti, G., Gordiyenko, Y., Santhanam, B., Wolf, J., Wiederhold, K., *et al.* (2017). Architecture of eukaryotic mRNA 3'-end processing machinery. Science *358*, 1056–1059.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). *6*, 80–92.

Corden, J.L. (1990). Tails of RNA polymerase II. Trends Biochem. Sci. *15*, 383–387.

Coulombe, C., Poitras, C., Nordell-Markovits, A., Brunelle, M., Lavoie, M.-A., Robert,

F., and Jacques, P.-É. (2014). VAP: a versatile aggregate profiler for efficient genome-wide data representation and discovery. Nucleic Acids Res. *42*, W485-93.

Creamer, T.J., Darby, M.M., Jamonnak, N., Schaughency, P., Hao, H., Wheelan, S.J., and Corden, J.L. (2011). Transcriptome-Wide Binding Sites for Components of the *Saccharomyces cerevisiae* Non-Poly(A) Termination Pathway: Nrd1, Nab3, and Sen1. PLoS Genet. *7*, e1002329.

Deshaware, S., and Singhal, R. (2017). Genetic variation in bitter taste receptor gene TAS2R38 , PROP taster status and their association with body mass index and food preferences in Indian population. Gene *627*, 363–368.

Fong, N., Brannan, K., Erickson, B., Kim, H., Cortazar, M.A., Sheridan, R.M., Nguyen, T., Karp, S., and Bentley, D.L. (2015). Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. Mol. Cell *60*, 256–267.

Gibbs, R.A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., Kovar, C., Lee, S., Muzny, D., Reid, J.G., Zhu, Y., *et al.* (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Gudipati, R.K., Villa, T., Boulay, J., and Libri, D. (2008). Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. Nat. Struct. Mol. Biol. *15*, 786–794.

Hirose, Y., and Ohkuma, Y. (2007). Phosphorylation of the C-terminal Domain of RNA Polymerase II Plays Central Roles in the Integrated Events of Eucaryotic Gene Expression. J. Biochem. *141*, 601–608.

Jensen, T.H., Jacquier, A., and Libri, D. (2013). Dealing with Pervasive Transcription. Mol. Cell *52*, 473–484.

Jeronimo, C., Watanabe, S., Kaplan, C.D., Peterson, C.L., and Robert, F. (2015). The Histone Chaperones FACT and Spt6 Restrict H2A.Z from Intragenic Locations. Mol. Cell *58*, 1113–1123.

Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010). Gene-specific RNA polymerase II phosphorylation and the CTD code. Nat. Struct. Mol. Biol. *17*, 1279–1286.

Kim, M., Krogan, N.J., Vasiljeva, L., Rando, O.J., Nedea, E., Greenblatt, J.F., and Buratowski, S. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature *432*, 517–522.

Ku, C.-S., Naidoo, N., and Pawitan, Y. (2011). Revisiting Mendelian disorders through

exome sequencing. Hum. Genet. *129*, 351–370.

Laitem, C., Zaborowska, J., Isa, N.F., Kufs, J., Dienstbier, M., and Murphy, S. (2015). CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II–transcribed genes. Nat. Struct. Mol. Biol. *22*, 396–403.

Lalonde, E., Albrecht, S., Ha, K.C.H., Jacob, K., Bolduc, N., Polychronakos, C., Dechelotte, P., Majewski, J., and Jabado, N. (2010). Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. Hum. Mutat. *31*, 918–923.

Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., *et al.* (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lek, M., Karczewski, K.J., Minikel, E. V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., *et al.* (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Lemay, J.-F., Marguerat, S., Larochelle, M., Liu, X., van Nues, R., Hunyadkürti, J., Hoque, M., Tian, B., Granneman, S., Bähler, J., *et al.* (2016). The Nrd1-like protein Seb1 coordinates cotranscriptional 3' end processing and polyadenylation site selection. Genes Dev. *30*, 1558–1572.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics *26*, 589–595.

Liang, K., and Keleş, S. (2012). Normalization of ChIP-seq data with control. BMC Bioinformatics *13*, 199.

Lindell, T.J., Weinberg, F., Morris, P.W., Roeder, R.G., and Rutter, W.J. (1970). Specific inhibition of nuclear RNA polymerase II by alpha-amanitin. Science *170*, 447–449.

Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., and Young, R.A. (2012). Revisiting global gene expression analysis. Cell *151*, 476–482.

Lunde, B.M., Reichow, S.L., Kim, M., Suh, H., Leeper, T.C., Yang, F., Mutschler, H.,

Buratowski, S., Meinhart, A., and Varani, G. (2010). Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. Nat. Struct. Mol. Biol. *17*, 1195–1201.

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Söding, J., and Cramer, P. (2010). Uniform transitions of the general RNA polymerase II transcription complex. Nat. Struct. Mol. Biol. *17*, 1272–1278.

Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., and Cramer, P. (2012). CTD Tyrosine Phosphorylation Impairs Termination Factor Recruitment to RNA Polymerase II. Science (80-.). *336*, 1723–1725.

Meinhart, A., and Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. Nature *430*, 223–226.

Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., *et al.* (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat. Genet. *42*, 790–793.

Orlando, D.A., Chen, M.W., Brown, V.E., Solanki, S., Choi, Y.J., Olson, E.R., Fritz, C.C., Bradner, J.E., and Guenther, M.G. (2014). Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome. Cell Rep. *9*, 1163–1170.

van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D., and Holstege, F.C.P. (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Rep. *4*, 387–393.

Porrua, O., and Libri, D. (2013). A bacterial-like mechanism for transcription termination by the Sen1p helicase in budding yeast. Nat. Struct. Mol. Biol. *20*, 884–891.

Porrua, O., and Libri, D. (2015). Transcription termination and the control of the transcriptome: why, where and how to stop. Nat. Rev. Mol. Cell Biol. *16*, 190–202.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Schreieck, A., Easter, A.D., Etzold, S., Wiederhold, K., Lidschreiber, M., Cramer, P., and Passmore, L.A. (2014). RNA polymerase II termination involves C-terminal-domain tyrosine dephosphorylation by CPF subunit Glc7. Nat. Struct. Mol. Biol. *21*, 175–179.

Shearwin, K.E., Callen, B.P., and Egan, J.B. (2005). Transcriptional interference--a crash course. Trends Genet. *21*, 339–345.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and

Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Suh, H., Ficarro, S.B., Kang, U.-B., Chun, Y., Marto, J.A., and Buratowski, S. (2016). Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II. Mol. Cell *61*, 297–304.

Taslim, C., Wu, J., Yan, P., Singer, G., Parvin, J., Huang, T., Lin, S., and Huang, K. (2009). Comparative study on ChIP-seq data: normalization and binding pattern characterization. Bioinformatics *25*, 2334–2340.

Thuriot, F., Buote, C., Gravel, E., Chénier, S., Désilets, V., Maranda, B., Waters, P.J., Jacques, P.-E., and Lévesque, S. (2018). Clinical validity of phenotype-driven analysis software PhenoVar as a diagnostic aid for clinical geneticists in the interpretation of whole-exome sequencing data. Genet Med.

Trakadis, Y.J., Buote, C., Therriault, J.-F., Jacques, P.-É., Larochelle, H., and Lévesque, S. (2014). PhenoVar: a phenotype-driven approach in clinical genomics for the diagnosis of polymalformative syndromes. BMC Med. Genomics *7*, 22.

Tudek, A., Porrua, O., Kabzinski, T., Lidschreiber, M., Kubicek, K., Fortova, A., Lacroute, F., Vanacova, S., Cramer, P., Stefl, R., *et al.* (2014). Molecular Basis for Coordinating Transcription Termination with Noncoding RNA Degradation. Mol. Cell *55*, 467–481.

Vannini, A., and Cramer, P. (2012). Conservation between the RNA Polymerase I, II, and III Transcription Initiation Machineries. Mol. Cell *45*, 439–446.

Vasiljeva, L., and Buratowski, S. (2006). Nrd1 Interacts with the Nuclear Exosome for 3′ Processing of RNA Polymerase II Transcripts. Mol. Cell *21*, 239–248.

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1–Nab3–Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. Nat. Struct. Mol. Biol. *15*, 795–804.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., *et al.* (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Weinmann, R., and Roeder, R.G. (1974). Role of DNA-dependent RNA polymerase 3 in the transcription of the tRNA and 5S RNA genes. Proc. Natl. Acad. Sci. U. S. A. *71*, 1790–1794.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5′ → 3′ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522–525.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zylber, E.A., and Penman, S. (1971). Products of RNA polymerases in HeLa cell nuclei. Proc. Natl. Acad. Sci. U. S. A. *68*, 2861–2865.